FDNY FIREFIGHTER TEST DEVELOPMENT AND VALIDATION REPORT

September 21, 2012

Submitted By:

PSI Services LLC

2950 North Hollywood Way, Suite 200
Burbank, CA 91505
(818) 847-6180 • (800) 367-1565

CONTENTS

TABLES

EXECUTIVE SUMMARY

The New York City Fire Department (FDNY) provides fire protection and life safety services for over eight million residents of the City, as well as providing fire prevention inspection services. To provide these critical services, the City employs over 8,200 Firefighters located in approximately 215 firehouses throughout the City. In light of the crucial role of Firefighters, it is imperative that the City recruit and select entry-level candidates who will be successful in learning in both the academy and performing the Firefighter role. To this end, the City retained PSI Services LLC (PSI) in 2010 to design and develop a job-related assessment instrument for selecting Firefighter recruits.

The project was undertaken in the context of ongoing litigation, in which the City of New York was found liable for hiring discrimination against Black and Hispanic applicants in two prior open-competitive examinations for the FDNY entry-level Firefighter position. Accordingly, the Court ordered the City of New York, in conjunction with the other parties to the litigation, to develop a new selection device for that position that would comply with laws against employment discrimination and meet professional standards of job-relatedness and business necessity, while minimizing adverse impact against racial/ethnic subgroups.

The test development and validation project was completed by PSI, in consultation with testing experts representing the U.S. Department of Justice (DOJ) and the Vulcan Society. In addition, a testing expert representing the court-appointed Special Master, Mary Jo White, attended all expert meetings and reviewed the work as it progressed.

Project Scope

The purpose of the project was to develop a lawful selection instrument for FDNY Firefighters that is job-related and valid, fair and objective, and practical and efficient for the City to administer and score on a large scale to tens of thousands of candidates. The instrument was designed to serve as an initial step in the selection process, providing a job-related means for selecting candidates to participate in subsequent steps of the hiring process (e.g., physical ability test, background investigation, and medical and psychological examination).

The project was designed and conducted in cognizance of professional testing standards and principles, including the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999) and the *Principles for the Validation and Use of Selection Procedures* (SIOP, 2003). Furthermore, the project was conducted to comply with the *Uniform Guidelines on Employee Selection Procedures* (EEOC et al, 1978). To this end, the project included:

- A job analysis study of the FDNY Firefighter position, which identified important abilities and characteristics that are essential to successfully learn and perform important FDNY Firefighter job tasks, and which would be appropriate and feasible to assess in the new selection test. Factors that are central to evaluating both probationary academy and on-the-job performance of FDNY Firefighters were also identified.

- Development of a new computer-based test (CBT) incorporating multimedia test item types to enable simulation of the manner in which newly hired Firefighters learn in the probationary fire academy and on the job. A noncognitive component was also developed

to assess personal characteristics that are important for successful Firefighter performance.

- Three validation strategies and studies conducted with several hundred FDNY Firefighters to gather empirical data that would enable construction of a job-related test and scoring method, and to yield validity evidence supporting the use of test scores in selecting candidates to advance in the Firefighter selection process.

- Development of a test scoring method prior to the administration of the test to candidates. After the candidates were tested, certain item scoring modifications were determined by a Test Validation Board (TVB), which was empanelled pursuant to New York law and court order.

Firefighter Job Analysis

PSI conducted a job analysis study of the FDNY Firefighter position to serve as the basis for determining appropriate content for a new Firefighter test. The objective of the study was to identify the most important abilities and characteristics required at the time of hire for successful performance of the FDNY Firefighter job. In conducting the study, particular attention was paid to ensuring that the job analysis was inclusive of various groups within FDNY (e.g., boroughs, engine and ladder companies, Firefighters and supervisors, and racial/ethnic and gender groups).

The job analysis entailed a combination of data gathering methods, including reviewing previous job analysis studies undertaken by the City and by PSI in other jurisdictions, job observation, focus groups and surveys. Testing consultants conducted job observations and interviews at two firehouses that each included engine and ladder companies, and at the FDNY training academy. A focus group was then convened with a diverse panel of 14 Firefighters and Officers, who reviewed draft lists of Firefighter job tasks, abilities and characteristics, along with draft instructions and rating scales for use in job analysis survey instruments.

Job analysis survey instruments were developed to elicit Firefighters' and Officers' ratings of the relative importance and frequency of performance/application of various job tasks, abilities and characteristics. A diverse sample of FDNY Firefighters and Officers was selected by PSI to attend a workshop at the FDNY academy to complete the job analysis surveys. Over 400 Firefighters and Officers attended the sessions and completed the surveys. The survey ratings were entered into an electronic database and, after quality control review, statistical analyses were conducted to identify core tasks and abilities/characteristics; i.e., those rated as performed or used by at least a two-thirds majority of Firefighters and Officers, and which were rated as important, very important or critical, on average. The surveyed Firefighters and Officers indicated that over 90% of the FDNY Firefighter job was covered in the survey, on average.

Next, a diverse panel of 37 Firefighters and Officers was convened at the FDNY academy to link important Firefighter abilities and characteristics to the core tasks, thereby identifying abilities and characteristics that are essential for the performance of the core tasks (that had been grouped into 17 Task Categories). As a result of this process, 49 core Firefighter abilities and characteristics were identified (i.e., rated as essential to performing a Task Category by at least 2/3 of the Firefighters and Officers), including: 18 cognitive abilities (reading, listening, learning, etc.); 24 noncognitive characteristics (willingness to adapt to a team; willingness to

accept responsibility, etc.); and 7 physical abilities (using muscular force, maintaining physical effort, etc.).

A supplemental analysis of Firefighter reading demands and learning protocols was conducted to gather additional job information for use in design of the test. The analysis of reading demands entailed compiling electronic copies of documents used by Firefighters in the probationary academy and on the job. A reading level (readability) analysis of the documents (totaling 4,282 pages) was conducted to gauge the relative difficulty of the materials and provide a benchmark for the CBT content related to reading comprehension.

In addition, the testing consultants conducted individual structured telephone interviews with ten firefighters and trainers to collect information regarding the use of reading materials in academy training and on the job. They were also asked how basic arithmetic is applied in training and on the job. The findings indicated that all Firefighters interviewed spent time reading the probationary training manual and reading on the job; most respondents indicated that information in manuals, reading lists, and other materials was also addressed with other learning sources (e.g., lectures, demonstrations, and videos). All interviewees confirmed that Firefighters must apply basic arithmetic skills without the use of a calculator.

The job analysis results also served as a basis for developing a Firefighter job performance measure that would later be used in a criterion-related validation study of the new test. A research-only Job Performance Rating Booklet was developed specifically for use in the validation study as a means for Lieutenants and Captains to evaluate the job performance of Firefighter incumbents. The rating booklet was developed with the input and review of Firefighters and Officers in a focus group session. The resulting rating booklet contained instructions, rating scales and job performance rating dimensions that covered 18 work areas, which represented core tasks and abilities/characteristics identified in the job analysis. Rating scales for overall effectiveness, relative ranking compared to other Firefighters, and work outcomes were also included.

Development of the Firefighter CBT

The CBT was designed specifically to measure cognitive abilities and noncognitive characteristics identified in the FDNY Firefighter job analysis as essential for successful performance of Firefighter job tasks, and required upon entry into the probationary Firefighter Academy. The CBT does not address physical abilities, which will be assessed in a separate testing instrument.[1]

Test specifications were developed outlining content areas (assessment dimensions) to be measured and to guide development of test items. The assessment dimensions were developed by grouping the core abilities and characteristics identified in the FDNY Firefighter job analysis. PSI, DOJ and Vulcan Society testing experts reviewed the abilities and characteristics and classified them into ten assessment dimensions, linking them to literature-based constructs, as follows:

[1] The Candidate Physical Ability Test (CPAT) is presently used by the City to assess Firefighter candidate physical capabilities.

- Cognitive abilities: Reading Comprehension, Ability to Learn and Apply Information, Listening Comprehension, Reasoning, and Basic Arithmetic; and

- Noncognitive characteristics: Conscientiousness, Agreeableness, Emotional Stability, Interpersonal Competence, and Honesty/Socialization.

Specifications for test item development were developed giving consideration to formats and item types that were: (a) realistic and similar to firefighter training and on the job, incorporating multiple modes of information (reading, lecture, demonstration, and observation); (b) not heavily reliant upon reading comprehension for assessment dimensions where reading was not the targeted ability; (c) reading comprehension passages calibrated to a reading level comparable to materials used by FDNY Firefighters; (d) reliable, objective, practical and amenable to automated delivery and scoring with a candidate pool numbering in the tens of thousands; and (e) resistant to cheating and harvesting (stealing) of content.

The resulting test specifications for cognitive assessment dimensions included a video-based simulation of an entry-level Firefighter's initial assignment to the FDNY training academy, supplemented by written excerpts from a simulated operations manual. Specifications representing the noncognitive assessment dimensions included traditional multiple-choice questions of several types (biographical, work attitude, and work-related personality); 14 noncognitive scales were developed to represent facets of the noncognitive assessment dimensions. Computer-based test (CBT) delivery of test questions was specified to support controlled delivery of multimedia question types that incorporate video, audio and graphic images. The CBT would enable secure delivery of alternate test forms and scrambling of test questions, which would help minimize cheating and content harvesting.

PSI drafted scenario scripts and test items referencing fictitious equipment to create a training simulation (instructor delivering lessons to a student) to provide a unified theme within which to address the cognitive assessment dimensions. Noncognitive items were written using established formats and question themes representing the various core Firefighter characteristics. Enactments of the scenarios were video-taped and the videos and test items were reviewed by testing experts representing PSI, DOJ, and the Vulcan Society; edits and enhancements were made as a result.

An experimental version of the Firefighter CBT was assembled for use in validation studies. The experimental version contained extra test items to allow for the subsequent selection of test item sets that best related to job performance while minimizing subgroup differences.

The final validated version of the Firefighter CBT (Form A) was comprised of a 57-item cognitive portion, which included a 3-part video lesson and simulated operations manual; and a 65-item noncognitive portion (background questions), which included six scales. A Training Guide was developed for use by examinees during the CBT to briefly study before each video lesson, to take notes during the video lesson, and to refer to in answering the test questions. The total examination time was four hours, including time for a tutorial and practice questions.

In light of the large number of job applicants (approximately 60,000) who signed up to take the Firefighter test, additional alternate CBT forms were developed to help safeguard the security and integrity of the examination; i.e., to minimize the exposure of the test items to candidates and create a barrier to cheating. The alternate forms contained unique cognitive portions and

used the same noncognitive items. The rationale for using the same noncognitive items on the alternate forms was that they are difficult to clone and are not subject to the same types of cheating concerns as cognitive items.

To ensure that the alternate CBT forms would be equivalent to the validated CBT Form A, items were developed following a cloning process wherein the Form A training lesson scripts were modified to create three alternate versions (B, C and D). The same item types were carried forward in the cloned item to match the original item type. The reading levels of the passages on the alternate forms were calibrated to the same level as Form A across alternate forms. An alternate version of the Training Guide was developed for each alternate form, again by cloning the original version used in Form A. The alternate CBT forms were reviewed and edited by testing experts representing PSI, DOJ, and the Vulcan Society. Extra items were written for each set of video training lessons which enabled the creation of two overlapping versions of each alternate form, resulting in six forms (B1, B2, C1, C2, D1 and D2). Two of the alternate forms (C2 and D2) were held in reserve and not administered to candidates.

Validity Evidence

Three strategies were followed to establish validity evidence for the new Firefighter CBT, which comply with the *Uniform Guidelines*: (1) content validity (Sec. 14C); (2) criterion-related validity (Sec. 14B); and (3) construct validity (Sec. 14D). An additional empirical study was conducted to assemble alternate forms of the CBT and to document their equivalence to the CBT (Form A) that was validated in the below described studies.

Content Validity

Content-oriented validity evidence for the cognitive portion of the Firefighter CBT was compiled in a study that was designed to establish a linkage between the content of the test and abilities that are essential for the performance of the FDNY Firefighter job. The study entailed selection of a diverse sample of 29 FDNY Firefighters and Officers to serve as job experts at a content validation workshop in July 2011.

During the workshop, the job experts first completed the cognitive portion of the CBT (Form A) to ensure that they had a complete understanding of the test. Then the participants were assembled as a group and instructed to complete a rating form to indicate the extent to which each of the 18 core cognitive abilities identified in the FDNY Firefighter job analysis was important to successfully complete the exercises contained in the CBT. The session also afforded an opportunity for the Firefighters and Officers to pilot the CBT to ensure that the instructions and testing procedures were clear, and that the CBT system was functioning properly in the DCAS computer-testing center.

The content validation procedure yielded results supporting the job relevance of the new Firefighter test; 17 of the 18 core Firefighter cognitive abilities were rated by a 2/3 majority of Firefighters and Officers as important or critical for successful performance one or more of the test exercises that were included in the final CBT Form A.

Criterion-Related Validity

Evidence of criterion-related validity for the Firefighter CBT was demonstrated in an empirical study that yielded statistically significant correlations between Firefighters' CBT scores and their performance both in the probationary academy and on the job.

A sample of 811 FDNY Firefighters was selected by PSI to complete CBT Form A, and for whom job performance data would be collected. The sample focused exclusively on full-time Firefighters who were assigned to ladder or engine companies in the five boroughs of New York. The sample was designed to be large enough to support statistical analyses of validity and fairness for racial/ethnic and gender groups. The participants were randomly selected within sampling strata, with over-representation of some groups to facilitate statistical analyses.

PSI worked with the FDNY to identify the selected Firefighters' supervisors, who would be asked to attend a special session to complete a confidential research-only performance evaluation (the afore-mentioned Job Performance Rating Booklet). For a subset of Firefighters, a second evaluator was identified to complete a performance evaluation to enable the computation of inter-rater reliability estimates for the performance ratings. The department identified over 400 Fire Lieutenants and Captains and scheduled them to attend one of 11 PSI-conducted rating sessions at the FDNY Training Academy in June 2011. During the session, participants were provided an orientation to the project, instructions for how to complete the ratings, and a facilitated discussion of examples of good and poor performance before rating the Firefighters.

In addition, PSI obtained the same Firefighters' prior training records from the FDNY academy during June and July of 2011. These data included academy test scores (quizzes, midterm exam, final exam) and practical exercise scores obtained by academy classes held between 2000 and 2008.

An electronic database was constructed and preliminary analyses of the job performance ratings and academy data were conducted to ensure the quality of the data for purposes of conducting the criterion-related validation study. Useable job performance ratings were obtained for 755 Firefighters, and useable academy data were obtained for 598 Firefighters.

In September 2011, the selected Firefighters completed the experimental CBT (Form A) at one of 16 testing sessions conducted at DCAS testing centers in Manhattan and Brooklyn. The Firefighters received advance letters from the FDNY and the union describing the project, its importance, and the need for confidentiality of the test materials. The testing sessions were supervised by PSI and were proctored by PSI and DCAS staff. During the test, proctors actively walked around the testing room to ensure that no one was talking, cheating or attempting to record the test content. The Firefighters completed the CBT at an individual pace and when finished, checked-out with the proctor who collected all notebooks and instructions. There were no incidents of unusual or irregular behavior reported by the test proctors, and no incidents of CBT technology failure or anomalies reported. A total of 735 Firefighters participated in the testing sessions.

Next, the Firefighter test data were extracted from the CBT system to create an electronic file for purposes of analysis. The test data were subjected to quality control steps and cases were

excluded if they failed to meet criteria for reasonableness (e.g., omitted too many items; did not spend a reasonable amount of time; scored below chance/random guessing level). The resulting sample sizes for the various portions of the test ranged from 682 to 718; and 612 Firefighters had complete data across all test portions.

A series of statistical analyses were conducted examining psychometric properties of the test items and test components. The purpose of the analysis was to identify and eliminate any poorly functioning items, and to examine the statistical properties of the test components.

Criterion-related validity of the CBT was then examined. Product-moment correlation coefficients were computed between Firefighters' scores on various parts of the experimental CBT and the measures of job performance and academy performance to examine the criterion-related validity of the test. The experimental CBT was divided into 10 cognitive components and 14 noncognitive components (or scales). All of the cognitive test components were significantly correlated with performance in the Firefighter academy (significant correlations ranged from .09 to .40, corrected for criterion unreliability);[2] and nine test components predicted at least one job performance composite (significant correlations ranged from .11 to .26, corrected for criterion unreliably).[3] Seven of the noncognitive components predicted at least one job performance rating (significant correlations ranged from .11 to .26, corrected for criterion unreliability).

The final CBT (Form A) was assembled in consideration of several factors, including: (a) maximizing validity; (b) representing important test content areas that were linked to core Firefighter abilities and characteristics; (c) minimizing racial/ethnic and gender group score differences; (d) yielding reliable scores with sufficient variance to be useful as a selection tool, and (e) practical administration time.

The validity results for the cognitive test components were examined to identify a subset that would address the above considerations. The testing experts selected a 3-part video learning exercise and operations manual. Similarly, the experts reviewed the noncognitive scale validity results in conjunction with the core Firefighter characteristics and noncognitive assessment literature, and identified six scales for inclusion in the CBT: Agreeableness, Dependability, Even Tempered, Low Anxiety, Self Esteem, and Activity. Validity coefficients for the final CBT total score ranged from .24 (predicting Overall Job performance Ranking) to .30 (predicting Academy Performance), correcting for criterion unreliability. To better estimate the true validity of the CBT, an additional adjustment was applied for restriction in the range of test scores in the criterion-related study relative to the range of candidate scores, which resulted in validity coefficients for the final CBT ranging from .32 (predicting Overall Job Performance Ranking) to .39 (predicting Academy Performance).[4]

Construct Validity

Evidence of construct validity was examined by correlating Firefighters' scores on cognitive and noncognitive portions with their scores on other previously validated (marker) tests of the same

[2] The reliability of the Academy performance exam composite was assumed to be .80, per the meta-analytic value suggested by Hunter and Hunter (1984).
[3] The reliability of the performance ratings ranged from .21 to .37, based on the criterion-related study.
[4] The CBT total score SD for candidates was 15.94; the SD for Firefighters in the validation study was 11.74.

or similar abilities and characteristics (constructs) that were administered during the criterion-related validation study.

Correlations were examined between the CBT noncognitive scales and three well-established measures of personality characteristics, including the Big-Five Inventory (BFI; John, 1991), the California Personality Inventory (CPI; Gough, 1996); and the Personnel Reaction Blank (PRB; Gough, 1971). The results indicated that the six CBT scales selected for the final CBT were significantly correlated with corresponding personality test scores in a manner that would be expected.

Correlations were examined between CBT cognitive scores and four well-established tests of cognitive ability from the Employee Aptitude Survey (EAS) test series measuring verbal, reasoning, perceptual, and numerical abilities (Ruch, et al. 2001). Overall, the pattern of correlations reflected the integrated nature of the training simulation, as the correlations with marker tests did not isolate specific constructs.

Equivalency Study

A subsequent study was conducted to pretest new items, assemble equivalent forms and document their equivalence to the validated CBT Form A. The study was conducted during January and March 2012, in Los Angeles, CA.

The study design called for a sample of 675 people to complete an experimental version of one of the alternate CBT forms (B, C, or D), along with Form A. The sampling plan targeted people from a range of hourly blue-collar jobs (e.g., construction, light industrial, and medical technology), with a high school diploma or some college completed (not highly degreed). Equal numbers of participants were sought for the most prevalent racial/ethnic groups in New York City (Black, Hispanic and White) to enable subgroup analyses.

PSI engaged a staffing services firm to source participants for the study. Participants were paid and were offered an incentive to perform well on the test (entry into a gift card lottery). The staffing firm used focused recruiting to recruit subjects in the target demographic characteristics.

In January 2012, a total of 46 testing sessions were conducted at four PSI testing centers located in the Los Angeles area. There was no mention of NY Firefighters and all test materials were sanitized and referred to only as the "Vocational Test Study." The test sessions were conducted by PSI test proctors who ensured that the Firefighter CBT was administered in a secure and standardized manner. Participants were randomly assigned to a computer station and CBT Form A, plus one of four alternate forms in counter-balanced order. A total of 718 people were tested.

The equivalency study data were extracted from the CBT system to create an electronic file, and were subjected to a quality control review. Data records were excluded from the analysis if they obtained a total cognitive score that would result from random guessing. A total of 688 examinees had complete data on both Form A and one alternate form (B, C, D).

Similar to the procedure for assembling CBT Form A, statistical analyses were conducted examining psychometric properties of the cognitive test items to identify and eliminate any poorly functioning items, and to examine the properties of the various test portions for purposes

of assembling alternate forms of the validated CBT Form A. Three primary alternate forms were assembled (B1, C1 and D1) which had unique video-based lessons and training guides. Three additional forms (Forms B2, C2 and D2) were assembled which overlapped with the primary forms, using the remaining items with the same video lessons.

Overall, the alternate CBT Forms were found to be highly correlated with the validated Form A ($r >= .87$, uncorrected), indicating that the forms provide comparable measurement of candidate abilities and characteristics. The forms were equivalent with respect to item content, raw score means and reliability. The forms varied slightly with regard to standard deviations in the equivalency sample, indicating that a small scale adjustment would likely be required to locate examinee scores on exactly the same scale of measurement. This standardization was later conducted using the much larger candidate sample to ensure that the new forms of the test would meet professional standards for test equivalency.

Investigation of Fairness

An analysis was conducted to examine the relationship between test scores and job performance where technically feasible for racial/ethnic groups (Blacks and Hispanics) to ensure the tests were fair; i.e., did not substantially under-predict actual job performance. The analysis was consistent with the definition of fairness set forth in the *Uniform Guidelines* [Sec. 14.B. (8) (a)].

To this end, a linear regression equation was derived for CBT (Form A) scores to predict job performance ratings for the Firefighters in the criterion-related validation study. Firefighters' predicted job performance ratings were compared to their actual ratings to compute a residual job performance score for Blacks, Hispanics and Whites (residual=actual minus predicted job performance rating).

Results of the fairness analysis indicated that the CBT does not unfairly predict minority groups' job performance. This finding is consistent with generally reported findings in the scientific testing literature (SIOP, 2003, p. 32).

Scoring and Use of the Firefighter Test

Scoring procedures for the CBT were developed by PSI prior to its administration to candidates. After the test was administered, a Test Validation Board (described below) identified scoring key modifications for certain items and these were incorporated into the final scoring, along with mandated credits (bonus points).

The scoring procedure included: (a) the assignment of weights to different portions of the CBT (weights determine the relative contribution of a portion to total CBT score); (b) application of a minimum passing score; (c) converting CBT scores to a 100-point integer scale; and (d) addition of mandated special credits for veterans, New York City residents, and candidates whose parent or sibling was a Firefighter killed in the line of duty.

Scoring weights were derived within the CBT cognitive test portion, giving 70% weight to the video lesson and 30% weight to the operations manual. These weights reflect the relative importance of the tasks to which the abilities tested are linked, scaled to a percentage.

Scoring weights were also derived for combining scores on the cognitive and noncognitive portions of the CBT, giving equal (50%) weight to the cognitive and noncognitive portions of the CBT. This 50/50 weighing approach was near the mid-point between two alternatives that were examined: (1) the job analysis-based approach, which gave more weight to noncognitive, and (2) a multiple regression approach, where regression weights were derived for cognitive and noncognitive portion scores in predicting a job performance rating composite, and which gave more weight to cognitive.

A minimum passing score (cut score) was derived for the CBT drawing from the criterion-related validation study data. A statistical analysis (linear regression) was conducted to identify the CBT score that corresponded to an average job performance rating reflecting minimally acceptable performance (a rating of "3-Just Adequate") on the job performance rating scale, adjusting downward for the standard error of estimate (SEE) of the regression equation.

CBT total scores were converted to a 100-point integer (whole number) score scale. The scaling procedure used two linear formulas: one for scores at or above the cut score to convert to a scale of 70 to 100; the other for scores below the cut score to convert to a scale of 0 to 69. Score conversion formulas were derived for each form of the CBT.

The score scale was developed in light of: (a) NY State and City requirements for setting the passing score to 70 and for adding bonus points to eligible candidates; (b) NY State and City requirements for adding bonus points to eligible candidates' scores; (c) the precision of measurement associated with CBT total scores (alpha=.88), which supported integer score units; and (d) criterion-related validity results supporting the use of CBT scores to select candidates who are more likely to be successful. To explore the reasonableness of the score scale prior to its use with candidates, it was applied to the Equivalency study sample. The scale was found to yield a fairly even distribution of scores that did not result in a large percentage of candidates obtaining the same score. Further, the 100-point scale enabled adding mandated bonus points in a straight-forward manner that could be readily explained to candidates.

Administration of the CBT

The CBT was administered to candidates between March 15, 2012 and August 1, 2012 in fifteen testing centers located throughout New York City and several outlying areas. A total of 42,231 candidates were tested, of whom 873 candidates were Emergency Medical Technicians and Paramedics that worked for the FDNY who took the exam on a promotional basis; and 41,358 were open competitive pool candidates. The City referred to the exam that the promotional candidates took as "Exam 2500" and the open competitive exam as "Exam 2000."

The testing sessions were proctored by trained PSI and DCAS staff that actively monitored candidates during the testing sessions. During check-in, candidates placed their cell phones and all belonging into a sealed plastic bag, which was then placed under their seats. Candidates were not allowed to talk during the test. Each computer station was separated by a carrel to prevent candidates from looking at other computers. The test items were delivered in random order within portions of the test and items were not numbered, so it would be extremely difficult for candidates to copy answers from other candidates.

Four forms of the CBT were administered to candidates. Forms B1, B2, and C1 were used during the first week. Form D1 was introduced on April 1. Forms C2 and D2 were not administered and were held in reserve in the event that additional forms were need in order to assure test security.

During administration of the exam, candidate performance on the test was monitored regularly and there was no evidence that scores increased over time, indicating that the exam content was not compromised.

Post-Administration Analyses

After the CBT was administered to candidates, an item analysis was conducted to confirm that the test items were properly scored. The results of the item analysis for each CBT form indicated that the test item key was appropriately applied and that the test items exhibited acceptable psychometric properties in the candidate sample (i.e., acceptable difficulty, positive item score correlation with total score, and distractor effectiveness).

A Test Validation Board (TVB) was convened, per New York State Civil Service Law and court order, during the month of June 2012 to review and adjudicate item protests that were submitted by candidates during formal test review sessions conducted after the test administration. The TVB recommended adjustments as follows: three cognitive items were identified for which one response option in addition to the originally keyed response would be counted as "full credit"; and six noncognitive items were identified for which full credit would be awarded for any of the response options. The TVB also awarded partial credit for six additional noncognitive items.

In addition to the TVB modifications, two cognitive test items on Form D were given full credit for all examinees because of a computer delivery issue that resulted in several hundred candidates receiving extraneous information with the delivery of those items. While it was not apparent that this caused a problem, the fair action to take was to effectively neutralize these two items by giving all examinees credit.

These item scoring modifications were enacted prior to computing candidates' final CBT scores. Subsequent analyses indicated that the scoring modifications did not affect the equivalence of the alternate forms.

CBT total scores were computed for candidates following the scoring procedure outlined earlier and summarized below. This procedure was applied separately for candidates taking each CBT Form (B1, B2, C1 and D1):

1. Compute CBT cognitive score, applying 70/30 weighting to the video learning and operation manual components;
2. Compute Total CBT score, applying 50/50 weighting to the cognitive and noncognitive portions, rounded to one decimal;
3. Apply minimum passing score ;
4. Convert scores to 100-point integer scale (rounded with no decimals), with passing score=70; and
5. Add bonus points to compute final CBT score.

The resulting candidate pass rate was 97.8% overall, 99.5% for promotional candidates and 97.7% for open pool candidates.

Use of CBT Scores to Select Candidates

The Firefighter CBT was developed and validated for use in assessing Firefighter candidates as a competitive examination (i.e., for use in determining continuation in the hiring process in descending order of score). DCAS has specified a manner of use of CBT scores in accordance with civil service law and its established regulations, as described below.

DCAS will create two separate eligible lists of passing candidates, one for the promotional candidates (Exam 2500), and one for the open competitive pool of candidates (Exam 2000). DCAS will exhaust the promotional candidate eligible list before selecting candidates from the open competitive pool. Both lists are top down rankings based on final CBT scores after applying all bonus points claimed by the candidates. The combination of final CBT score and bonus points is called the "Adjusted Final Score." All candidates will be assigned a list number for administrative purposes only based on the last five and then the first four digits of their social security numbers (SSNs) with the higher SSNs receiving the lower list numbers. This list number will be used, if necessary, to break ties among candidates with the same Adjusted Final Score

The FDNY appoints a class of approximately 300 probationary Firefighters at a time. In order to provide a final group of 300 candidates who successfully complete all remaining steps in the hiring process, the FDNY requests that DCAS certify 1200 names to the FDNY at a time. However, because all candidates with tied Adjusted Final Scores must be certified at the same time, there may be instances when more than 1200 candidates are certified to the FDNY at once. Typically, the FDNY will appoint two academy classes per year, and thus would hire 600 probationary Firefighters per year, requiring certification of approximately 2400 candidates from the eligible lists.

All candidates certified by DCAS to the FDNY are scheduled for CPAT (Candidate Physical Ability Test) orientation sessions (as required by the test license) and afforded the opportunity to train prior to administration of the CPAT. Selected candidates also receive employment packets from FDNY's Candidate Investigation Division (CID), which schedules an intake interview. The background investigation is then conducted and candidates undergo psychological and medical screening.

Projected Selection Rates over the Life of the Eligible Lists

Projected annual candidate selection rates were made based on: (a) the candidate score results; (b) the number of Firefighter trainee positions available on an annual basis (approximately 600); (c) DCAS' above described procedures for establishing eligible lists and selecting candidates; and (d) the number of candidates the City has historically selected from the eligible lists to yield a sufficient number to fill an academy class (a select/hire ratio of approximately 1.5 to 1 is typically needed for promotional candidates and about 4 to 1 for open competitive candidates).

A total of approximately 9,417 internal/promotional and open pool candidates (22.3%) are projected to be selected from the eligible lists over 4 years to fill 2400 positions. These estimates do not reflect the exact number of candidates who will pass the other steps in the selection

process (CPAT, background, medical and psychological exams) or the application of random (SSN-based) selection to break ties when more eligible candidates are available than positions.

Projected selection rates were computed within racial/ethnic and gender groups and analyses were conducted to identify potential adverse impact for the open competitive pool of candidates. These projected pass rates are subject to the same caveats noted above.

Potential adverse impact against racial/ethnic and gender subgroups was analyzed using the four-fifths rule (Uniform Guidelines Sec. 4.D) and application of a statistical significance test (the 2 or 3 standard deviation test). The four-fifths rule and statistical significance test were both satisfied over the full 4-year period that the City anticipates selecting Exam 2500 and Exam 2000 candidates from the eligible list (which is projected to be 9,417 candidates).

The minimum passing score used for both Exam 2000 and Exam 2500 also satisfied the four-fifths rule but it did not satisfy the statistical significance test with respect to Black internal candidates, and with respect to Black, Hispanic, Asian and Native American open pool candidates. The minimum passing score did not produce a significant disparate impact against men or women and, likewise, gender disparities in each of the four years that the City anticipates hiring were not significant.

In summary, there was no evidence that adverse impact would result in significant differences in the selection rates for any minority or female candidate groups from use of Final Adjusted Scores to select open pool candidates over the four-year life of the eligibility list.

CHAPTER 1: INTRODUCTION AND OVERVIEW OF THE PROJECT

Introduction

This report describes the methodology and results of an extensive research and development project undertaken for the City of New York to develop a new selection instrument for the entry-level Firefighter position in FDNY. This chapter provides background information about the project. Following chapters describe project steps that were completed to identify FDNY Firefighter job requirements and important capabilities and characteristics to measure on the new test; develop new test content; establish validity evidence for the test; develop a scoring method for the use of the new test to select candidates; and provide results and projected outcomes for probationary academy classes for the anticipated life of the list.

The project was conducted in cognizance of professional testing standards and principles, such as *the Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999) and the *Principles for the Validation and Use of Selection Procedures* (SIOP, 2003). Furthermore, the project was conducted to comply with the *Uniform Guidelines on Employee Selection Procedures* (EEOC et al, 1978). Accordingly, the report addresses requirements set forth in the *Uniform Guidelines,* Sec. 14 Technical Standards for Validity Studies, and Sec. 15 Requirements for Documentation of Impact and Validity Studies. Relevant sections of the *Uniform Guidelines* addressed by each section of the report are noted, where applicable.

Problem and Setting[5]

The New York City Fire Department (FDNY) provides fire protection and life safety services for over eight million residents of the city, as well as providing fire prevention inspection services. To provide these critical services, the city employs over 8,200 Firefighters located in approximately 215 firehouses throughout the city.

In light of the crucial role Firefighters play in the protection of life and property, it is imperative that the city recruits and selects entry-level candidates who will be successful in learning in the academy and performing in the Firefighter role. To this end, the city retained PSI Services LLC (PSI) in 2010 to design and develop a job-related assessment instrument for selecting Firefighter candidates, which would be valid as well as have no or minimal adverse impact.

It is important to note that this project was undertaken in the context of pending litigation, in which the City of New York was found liable for hiring discrimination against Black and Hispanic applicants for the FDNY entry-level Firefighter position. Accordingly, the Court ordered the City of New York, in conjunction with the other parties to the litigation, to develop a new lawful selection device for the position that, again, was valid and with attention to minimizing adverse impact.

The test development and validation project was completed by PSI, in consultation with testing experts representing the U.S. Department of Justice (DOJ) and the Vulcan Society. In addition, a testing expert representing the court-appointed Special Master attended all expert meetings and reviewed the work as it progressed. PSI's project team included John Weiner, Project Director;

[5]Uniform Guidelines Sec. 15 B. (2)

Joseph Abraham, Ph.D., Project Manager; Sheldon Zedeck, Ph.D., Project Advisor; and Donna Denning, Ph.D., Project Consultant. The DOJ project team included David Jones, Ph.D., Leaetta Hough, Ph.D., and Rand Gottschalk, M.A. Harold Goldstein, Ph.D., represented the Vulcan Society, and Shane Pittman, Ph.D., represented the court-appointed Special Master. The PSI, DOJ, and Vulcan Society teams met regularly in person and via teleconference and web meetings throughout the project to review, discuss and attempt to reach consensus on the project steps, methods and results.

Project Scope

The overarching purpose of the project was to develop a lawful selection procedure for FDNY Firefighters that is job-related and valid, fair and objective, and practical and efficient for the City to administer and score on a large scale with tens of thousands of candidates. To this end, the project included several extensive research and development components, including:

- A job analysis study of the FDNY Firefighter position, which identified important abilities and characteristics that are essential to successfully learn and perform important FDNY Firefighter job tasks, and which would be appropriate and feasible to assess in the new selection test. Factors that are central to evaluating both probationary academy and on-the-job performance of FDNY Firefighters were also identified.

- Development of a new computer-based test (CBT) incorporating multimedia test item types to enable simulation of the manner in which newly hired Firefighters learn in the probationary fire academy and on the job. A noncognitive component was also developed to assess personal characteristics that are important for successful Firefighter performance.

- Three validation strategies and studies conducted with several hundred FDNY Firefighters to gather empirical data to enable construction of a job-related test and scoring method, and to yield validity evidence supporting the use of test scores in selecting candidates to advance in the Firefighter selection process.

- Development of a test scoring method prior to the administration of the test to candidates. After the candidates were tested, certain item scoring modifications were determined by a Test Validation Board (TVB), which was empanelled pursuant to New York law and court order.

User(s), Locations(s) and Date(s) of Study[6]

User: New York Department of Citywide Administrative Services (DCAS)
Location: New York, NY
Dates of Study: July 10, 2010 to September 2012.

Contact Person: John Weiner, Chief Science Officer, PSI Services LLC, 2950 N. Hollywood Way, Suite 200, Burbank CA 91505.

[6] Uniform Guidelines Sec. 15 B. (1)

CHAPTER 2: FDNY FIREFIGHTER JOB ANALYSIS

Introduction

PSI conducted a comprehensive job analysis study of the FDNY Firefighter position to serve as a basis for determining appropriate content for a new Firefighter test. The primary goal of the job analysis was to identify the most important abilities and characteristics required at the time of hire (required "Day-1") for successful performance of the FDNY Firefighter job. The study also examined reading demands and the processes by which FDNY Firefighters acquire information in the academy and on the job. In conducting the study, particular attention was paid to ensuring that the job analysis was inclusive of various groups within FDNY (e.g., boroughs, engine/ladder companies, incumbents/supervisors, and racial/ethnic and gender groups).[7]

Job Titles and Codes[8]

The City's job title for the job in question and the corresponding job title(s) and code(s) from U.S. Department of Labor's online Occupational Network (O*NET) are as follows:

City Job Title	Department of Labor Job Code and Title
Firefighter	33-2011.00 – Firefighters 33-2011.01 – Municipal Firefighters

Job Data Collection: Observation, Focus Group and Surveys

Information about the FDNY Firefighter job was collected between September and December 2010, using a combination of data gathering methods, including reviewing previous job analysis studies undertaken by the City and by PSI in other jurisdictions, job observation, focus groups and surveys. An overview of the job analysis data collection procedure is provided below.

Documentation Review and Assembly of Preliminary Job Descriptive Information

As an initial step, PSI examined a variety of documents, including previous job analysis reports (DCAS 2007, 2002), the Probationary Firefighter Manual, training evaluation forms, and training books. In addition, PSI reviewed documentation from its previous consulting work with other jurisdictions; important Firefighter tasks and abilities/characteristics from previous PSI analyses; and job analysis documentation of the Firefighter job published by the U.S. Department of Labor (O*NET).

On the basis of this review, PSI assembled preliminary lists of potentially relevant Firefighter job tasks, abilities and work-related characteristics. The abilities and characteristics were defined using a job-specific behavioral approach (Goldstein, Zedeck, and Schneider, 1993), incorporating terminology that would be readily understood by Firefighters who served in later stages of the project as subject matter expert (SME) reviewers. Knowledge and skills typically

[7] Uniform Guidelines Section 14 A., B.(2)(4), C.(2), D.(2); Section 15 B. (3)(6), C.(3), D. (4)
[8] Uniform Guidelines Section 15 B.(4), D.(5)

acquired during training (after hire) were not included in the lists since the focus of the project was on entry-level job requirements.

Observation of the Work Environment

On September 16, 2010, testing consultants conducted job observations and interviews at two firehouses that each included both engine and ladder companies (Brooklyn Engine 282 & Ladder Co. 148, and Manhattan Engine 24 & Ladder Co 5), and at the FDNY training academy. The job observations included observing the firehouse environment first-hand; interviewing Firefighters and officers during the observations; riding along with Firefighters in response to an alarm; observing daily drills; sitting-in on an informal training session; and observing the training environment.

Focus Group Session

PSI then convened a focus group at the FDNY academy on September 17, 2010, to review the preliminary lists of job tasks and abilities/characteristics, along with draft survey instructions and rating scales for use in composing a job analysis survey instrument. A total of 14 FDNY personnel participated as subject matter experts (SMEs), including 10 incumbent Firefighters and 4 supervisors (lieutenants and captains). The participants were required to be in good standing and have at least 2 years tenure on the job, and were selected by PSI to represent the five boroughs, race/ethnicity, gender, and job level. A breakdown of participants is as follows: *Borough*: Manhattan (3), Bronx (3), Staten Island (1), Brooklyn (3), Queens (4); *Race/ethnicity*: Asian (1), Black (2), Hispanic (4), White (7); *Gender:* Female (3), and Male (11).

The session began with an orientation to the project and workshop procedure. The SMEs were then guided through a formal review of the preliminary lists of tasks and abilities/characteristics, and the drafts of survey rating scales and instructions for the job analysis survey. During this discussion, items were projected onto a screen and SMEs rated whether each task or ability/characteristic is important for successful performance as an entry-level Firefighter (each indicated "yes," "yes with change," or "no"). Participants also commented on the rating scales and instructions. PSI made changes to the wording suggested by the consensus of SMEs. SMEs identified any job-specific knowledge or physical skills that candidates would not be expected to possess before hire (i.e., would develop in training) and any such items were removed from the list.

At the conclusion of the session, SMEs rated the overall percent of the FDNY Firefighter job they considered to be covered by the updated lists of tasks and abilities/characteristics. They also were provided an opportunity to add tasks and abilities/characteristics. The tasks were rated as covering between 75% and 100% of the job by all participants (mean rating = 92.4%), while the abilities and characteristics were rated as covering between 75% and 100% of the job by 92.9% (13/14) of the participants (mean rating = 91.9%).

Draft Job Analysis Surveys and Final Review

Next, two survey instruments were drafted: a Job Task Survey and a Survey of Abilities and Characteristics. These instruments were subjected to a final review by the testing consultants and SMEs.

The testing consultants, upon review of the ability/characteristic section list, added several items to ensure full representation of the abilities and characteristics from a psychometric standpoint. Additionally, for the purpose of ensuring data quality when the job analysis survey tools were administered, five "quality check" task items and five "quality check" ability/characteristic items were developed to identify individuals completing the survey who may have responded carelessly. For example, the task statement *"Submits daily staffing reports to office of staffing"* was added to the survey. Individuals who would rate this and other such items as being part of the Firefighter job were considered to have spent little attention on the task, and were excluded from the research sample. Appendix A lists the quality check survey items.

A group of four SMEs was convened to review the draft surveys via webinar on October 19, 2010; participants included one Firefighter, one lieutenant, and two battalion chiefs. As a result of the review, minor adjustments were made to the wording of several tasks and rating scales, as well as the organization of tasks into task groupings, or clusters.

Final Job Analysis Surveys

The final Job Task Survey and a Survey of Abilities and Characteristics are described below.

The Job Task Survey contained 187 tasks (plus 5 quality check items) and included three rating scales for participants to rate each task: Performance, i.e., whether or not the task is performed (Yes/No); Importance to the overall job of an entry-level Firefighter (1–5 scale); and the Frequency with which the task is performed (1–7 scale). The survey instructed participants to complete the ratings with respect to the work performed by entry-level Firefighters at their respective stations. For purposes of the survey, an entry-level Firefighter was defined as: "someone who has completed Probationary Firefighter School and has been assigned to an engine or ladder company, but has not taken on any specialized work." Figure 1 shows the rating scales that were used in the Job Task Survey. Appendix B, Figure B-1 shows the job task rating scales and Appendix C, Figure C-1 contains the entire Job Task Survey and instructions.

A second survey instrument was developed that listed 120 abilities and characteristics (plus 5 quality check items) and included rating scales for participants to rate each item on two scales: Importance to the overall job of an entry-level Firefighter (1-5 scale); and Required-on-Day-1; that is, whether a new Firefighter must bring the ability/characteristic with them when entering Probationary Firefighter School (Yes/No). Participants were instructed to complete the ratings with respect to the requirements of a new Firefighter at the time of hire, before beginning training. Appendix B, Figure B-2 shows the ability/characteristic rating scales and Appendix C, Figure C-2 contains the full content and instructions for the Survey of Abilities and Characteristics.

Survey Sampling Plan and Onsite Administration[9]

Job Analysis Survey Sampling Plan

A survey sampling plan was designed to ensure that Firefighters who participated in the job analysis survey represented important demographic groups in the FDNY workforce. PSI selected survey participants based on a random-stratified sampling approach. A primary list was generated, along with a list of alternates to be added to the data collection process if the primary participants were unable to participate. The alternate list fell within the same demographic categories. A total of 408 Firefighters and officers were selected to participate in the survey. Appendix D displays the target sample for the survey with respect to borough, job level, engine/ladder company, race/ethnicity, and gender.

Survey Administration

The job analysis surveys were administered to incumbent FDNY Firefighters and Officers in three survey sessions conducted by PSI at the FDNY Academy on October 25, 26, and 27, 2010. These onsite sessions provided a means to ensure that the survey would be completed accurately and with appropriate attention to instructions.[10]

During the session, participants were provided an overview of the project and instructions for completing the surveys to ensure that they fully understood the process and the survey information. PSI session facilitators presented the following information at the beginning of each session: Background information on the project and on the purpose of the meeting; description of the survey development process and content; review of instructions; and discussion of rating scales and process, including example items. It was emphasized to participants that they should not proceed with the survey unless they had a full understanding of the assignment, and that they should ask questions of the facilitators should such questions arise.

Analysis of Core Firefighter Tasks and Abilities/Characteristics

Quality Control and Decision Rules

Prior to analyzing the Job Analysis Survey ratings, PSI conducted a rating quality analysis to identify and potentially exclude participants who may have answered carelessly, did not understand or did not follow instructions, or exhibited little or no variance in their ratings (suggesting the rater did not distinguish between the tasks, abilities and characteristics with regard to their relative importance of frequency of performance/use). Before implementing these rules, a data key-entry and verification process took place, as well as a check of all score ranges for out of range values. The rules used to exclude data on the basis of questionable rating quality are summarized in Appendix E.

The survey quality control criteria resulted in the exclusion of 36 Job Task Surveys and 61 Ability/Characteristic Surveys. Because different exclusion rules were used for the task and

[9] Uniform Guidelines Section 14 B.(4), 15 B. (6)

[10] The session was attended by a DOJ testing consultant and DCAS staff, who assisted with the distribution and review of completed surveys.

ability/characteristic surveys, the final sample sizes for survey analysis were 379 for the task survey and 354 for the ability/characteristic survey. Appendix F, Tables F-1 through F-3 display the demographic characteristics of participants retained for the task survey and the ability/characteristic survey.

As an additional quality control step, the following rules were used to identify and exclude problematic data for specific tasks and abilities/characteristics due to failure to follow instructions on individual items:

- For tasks rated as "not performed," any corresponding Importance and Frequency ratings were excluded;
- For tasks where the performance rating was missing, the corresponding Importance and Frequency ratings were excluded;
- For abilities/characteristics rated as "not important," any corresponding Day-1 ratings were excluded;
- For abilities/characteristics where the importance rating was missing, the corresponding Day-1 rating was excluded.

Reliability of Job Analysis Survey Ratings

The reliability of the job analysis survey ratings was assessed by computing the intra-class correlation coefficient (ICC) for each task and ability/characteristic rating scale.[11] The purpose of this analysis was to quantify the extent of agreement among survey participants in using the above-described rating scales to describe the job tasks, abilities and characteristics with respect to how frequently they are performed or used, their relative importance , and whether they are required on Day-1, prior to starting the probationary academy. The reliability index ranges from 0.0 (no agreement) to 1.0 (perfect) agreement. The reliability of the mean of all raters per scale (the basis for determination of core tasks and abilities/ characteristics) was high, exceeding .99 for each rating scale. Thus, it may be concluded that the ratings provided by Firefighter participants were of sound quality for purposes of identifying core job requirements.

Descriptive Statistics

Characteristics of the job analysis survey sample are summarized in Table 1 separately for Firefighters and Officers with respect to borough, gender, and race/ethnicity. Overall, the survey sample was diverse and inclusive of demographic groups comprising the FDNY Firefighter workforce. Characteristics of the job analysis survey sample are summarized in Table 1 with respect to job title/level, borough command, gender, and race/ethnicity. A total of 402 respondents completed a task survey, ability/characteristic survey, or both; these included 284 Firefighters and 118 Officers, representing 3.4% and 5.9% of the respective FDNY workforce levels.

[11] ICC was computed using a two-way random effects model with absolute agreement definition (Nichols, 1998). Adjustments were made for missing data in this computation because individuals were instructed to skip Task Importance and Task Frequency ratings if a task was rated as "Not Performed;" other missing ratings were replaced with the mean of other respondents' ratings.

Table 1. Characteristics of the Job Analysis Survey Sample

	Workforce		Surveys Completed		
Firefighters	No.	% of Population	No.	% of Population	% of Sample
Total	8374	100.0%	284	3.4%	100.0%
Borough					
Bronx	1814	21.7%	47	2.6%	16.5%
Brooklyn	1902	22.7%	73	3.8%	25.7%
Manhattan	1489	17.8%	36	2.4%	12.7%
Queens	1997	23.8%	58	2.9%	20.4%
Staten Island	1132	13.5%	70	6.2%	24.6%
Borough not specified	40	0.5%	0	0.0%	0.0%
Gender					
Female	24	0.3%	8	33.3%	2.8%
Male	8350	99.7%	276	3.3%	97.2%
Ethnicity					
Asian	76	0.9%	4	5.3%	1.4%
Native American	5	0.1%	0	0.0%	0.0%
African American	312	3.7%	19	6.1%	6.7%
Hispanic	639	7.6%	31	4.9%	10.9%
White	7342	87.7%	230	3.1%	81.0%
Officers					
Total	2043	100.0%	118	5.8%	100.0%
Borough					
Bronx	420	20.6%	29	6.9%	24.6%
Brooklyn	465	22.8%	10	2.2%	8.5%
Manhattan	369	18.1%	37	10.0%	31.4%
Queens	470	23.0%	23	4.9%	19.5%
Staten Island	279	13.7%	19	6.8%	16.1%
Borough not specified	40	2.0%	0	0.0%	0.0%
Gender					
Female	3	0.1%	1	33.3%	0.8%
Male	2040	99.9%	117	5.7%	99.2%
Ethnicity					
Asian	1	0.0%	0	0.0%	0.0%
Native American	1	0.0%	1	100.0%	0.8%
African American	28	1.4%	7	25.0%	5.9%
Hispanic	62	3.0%	15	24.2%	12.7%
White	1951	95.5%	95	4.9%	80.5%

Note: Final analysis sample of participants who completed either the Job Task Survey or the Survey of Abilities and Characteristics.

The task and ability/characteristic ratings were averaged across the survey participants for each scale to yield summary statistics for each rating scale, including: task percent performed, mean frequency performed and mean importance; and ability/characteristic mean importance and

percent required Day-1. These statistics were used in subsequent analyses to identify "core" Firefighter tasks, abilities and characteristics, and are reported in Appendices G and H, respectively.

At the conclusion of the task survey, respondents were asked to estimate the extent to which the survey covered the tasks performed by entry-level Firefighters at their station, by indicating a percentage between 0 and 100. The mean percent coverage rating for the task survey was 91%, with a standard deviation of 8.8 (N=371). Similarly, at the conclusion of the ability/characteristic survey, respondents were asked to estimate the extent to which the ability/characteristic survey covered the abilities and characteristics needed by entry-level Firefighters at their station, by indicating a percentage between 0 and 100. The mean percent coverage rating for the ability/characteristic survey was 91%, with a standard deviation of 9.1 (N=340). These results suggest that both the task and ability/characteristic surveys were comprehensive in their coverage of entry-level Firefighter tasks and abilities/characteristics.

Core Tasks and Ability/Characteristics

Analyses of the job task and ability characteristic mean ratings were conducted to identify "core" Firefighter tasks and abilities/characteristics that reflect the most important and commonly performed components of the FDNY Firefighter job. To this end, the task ratings were evaluated to identify those rated as being performed by a strong majority (2/3) of Firefighters and rated as very important or critical; and the ability/characteristic ratings were evaluated to identify those rated by a strong majority of Firefighters as needed on Day-1 and rated very important or critical for job success.

Pre-established criteria were applied to identify core tasks and abilities/characteristics, as follows:

Tasks:

1. Performed by at least a 2/3 majority of Firefighters (66.7%); and
2. Mean Importance rating of 4.0 (Very Important) or higher on the 5-point scale.

Abilities/Characteristics

1. Mean Importance rating of 4.0 (Very Important) or higher; and
2. Required Day-1 by at least a 2/3 majority (66.7%).

As a result of applying these criteria, 108 tasks, falling within 17 task clusters, were identified as core, as were 70 of the abilities and characteristics. The resulting core tasks are listed within Task Categories in Appendix I, Table I-1; the core abilities and characteristics are listed in Appendix I, Table I-2.

As another check on the quality of the job analysis survey ratings, they were compared to the City's 2007 job analysis results. The results were found to be highly consistent with respect to the number of common items, mean ratings, and relative importance. A total of 166 tasks were common between the 2010 and 2007 surveys and the mean importance ratings were found to differ by only 0.01 point on average (on the 5-point importance rating scale). The correlation between the mean importance ratings of the 166 common tasks was found to be very high

(r=.95), further indicating a high degree of consistency between the present task analysis study and the 2007 study results.[12]

Linkage of Abilities/Characteristics and Tasks

On December 2, 2010, PSI conducted a Linkage Survey meeting with Firefighter and officer SMEs to further document the importance and relevance of the abilities/characteristics to the tasks performed by FDNY Firefighters. Participants in the session completed a survey instrument to formally record their expert judgments regarding the importance of the abilities/characteristics in the performance of Firefighter task clusters.

SME Sample

The SMEs were selected by PSI to represent FDNY Firefighters with respect to borough, job level, engine and ladder companies, race/ethnicity, and gender using a stratified random sampling approach. As with prior sampling steps, PSI provided FDNY with a primary list of invitees, as well as a list of alternates. It is important to note that operational requirements and the voluntary nature of participation resulted in the need to invite alternates to the session in some cases. The Linkage Rating session participant sample is described in Table 2.

Table 2. Characteristics of Linkage Participant Sample

Borough	No. Firefighters	No. Supervisors
Manhattan	7	2
Bronx	3	5
Staten Island	1	3
Brooklyn	2	3
Queens	6	5
Race/Ethnicity		
Asian/Native American	3	0
Black	4	5
Hispanic	4	6
White	8	7
Gender		
Female	7	0
Male	12	18
Total	19	18

Linkage Survey

The linkage survey contained the 108 core tasks, grouped into 17 Task Categories, which resulted from the analysis of the Job Task Analysis Survey ratings. Because the analysis of

[12] Direct comparison of the 2010 ability/characteristic ratings to the 2007 ratings was not feasible because different approaches were taken in the two studies; the former used 120 specific behavioral statements, and the latter used 46 general construct-oriented statements. .

ability/characteristic ratings was still underway at the time of the linkage session, the testing consultants chose to include nearly all of the ability/characteristics in the linkage process. Ability/characteristic items were included if they were rated as having an importance of at least 3.75 (where 3=Important and 4=Very Important), without regard to Required Day-1 ratings; 108 ability/characteristics met this criterion.

In the interest of reducing potential rater fatigue and to enhance the accuracy of the linkages made by SMEs, PSI divided the ability/characteristics in half and created two forms of the linkage survey (Form A and Form B), each with 54 ability/characteristics. Half the SMEs completed Form A, and half completed Form B.

During the session, the SMEs were split into four groups, two groups completing Form A, and two groups completing Form B. To ensure consistent delivery of linkage survey introduction and instructions, PSI facilitated a plenary session with all participants prior to splitting into the four groups. SMEs were instructed to rate each ability/characteristic with respect to its criticality for performing tasks in each of the task categories, using the following rating scale:

Criticality Rating

E = Essential - This ability or characteristic is essential to the successful performance of one or more tasks in this category.

H = Helpful - This ability or characteristic is helpful in performing one or more tasks in this category. Tasks in this category could be performed without this ability or characteristic, although it would be more difficult or time consuming.

N = Not Relevant - This ability or characteristic is not needed to perform any task in this category.

In addition to the ability/characteristic-Task Category linkage ratings, a supplemental section was included in the survey. In this section, SMEs were asked to distribute 100 points among each of the core 17 Task Categories in accordance with their perceived relative importance for successful performance of the entry-level FDNY Firefighter job as a whole.

Appendix J, Figure J-1 displays the Form A linkage survey (Form B was identical except for a different listing of ability/characteristics). In addition to the survey, SMEs were provided with a reference document that listed the Task categories and associated tasks. This reference document is displayed in Appendix J, Figure J-2.

Quality Control of Linkage Ratings

As with the job analysis survey, a rater quality analysis was conducted to ensure the quality of survey responses, using the following survey exclusion rules.

1. Did not follow instructions: Linkage surveys were to be excluded if 5% or more of ratings were missing. No cases were identified with a sufficient number of missing ratings to qualify for exclusion.

2. Exhibited little or no variance: Linkage surveys were excluded if the standard deviation (SD) of ratings across all linkage ratings for a given participant was extremely low (>2

SDs below the mean SD); this criterion resulted in the removal of one rater who completed Form A, and one rater who completed Form B.

Reliability of Linkage Ratings

The reliability of the linkage ratings was assessed by computing the ICC in a similar manner to the analysis of job analysis survey ratings. The ICC was computed for survey Form A and Form B, after adjusting the data to account for missing ratings. The reliability of the average of all raters was high, with ICC values of .82 and .91 for Forms A and B, respectively.

Linkage Rating Results

An ability/characteristic was considered to be "linked" to a Task Category if it was rated as essential for successful performance of the Task Category by a strong majority of raters (at least 66.7%). An ability/characteristic was required to be linked to at least one Task Category to be included in the final list of core abilities/characteristics considered for inclusion in the design of the new test. A total of 49 core ability/characteristics were identified as essential for the performance of tasks in at least one Task Category. The linkage rating results are reported in Appendix K; cells in the table with shading indicate that the percentage of the sample endorsing an item as essential was at least 66.7%.

The resulting final list of 49 core abilities/characteristics is reported in Table 3 (original survey ID numbers are shown for each item). These abilities and characteristics represent a range of domains, including cognitive ability, physical ability, and other noncognitive characteristics.

Table 3. Firefighter Core Abilities and Characteristics

Cognitive Abilities
001-Ability to read and interpret short messages written in English (for example, notes, log entries, teleprinter tickets).
002-Ability to read routine documents written in English (for example, bulletins, articles, notices, announcements) to keep apprised of current job-related information.
003-Ability to read and interpret technical materials written in English (for example, instructional manuals, operating manuals, and other official FDNY documents) to learn new information and/or update job knowledge.
005-Ability to write brief notes/statements in English (for example, fill in forms, log entries, take messages) legibly, completely, and accurately.
007-Ability to listen to, and understand information on how to perform a task or series of tasks from a trainer or others (for example, instructions from a commanding officer, training information on specific steps to follow in different situations).
008-Ability to understand information presented orally in English, both in person (for example, in a training session) and from a variety of communications devices (for example, radio, phone, intercom).
009-Ability to listen to and understand people in emergency situations (for example, people who are upset, frightened, confused).
010-Ability to state ideas clearly and concisely when speaking in English (for example, giving instructions, explaining procedures, providing technical information).
016-Ability to quickly and accurately compare letters, numbers, information and objects (for example, addresses,

names, radio codes) to determine if they are the same or different.
017-Ability to concentrate on the work to be performed in spite of distractions, keeping aware of one's surroundings (for example, during a highway accident).[13]
018-Ability to remain attentive while performing routine or repetitive tasks (for example, taking up hose line).[13]
021-Ability to observe another person performing/demonstrating an activity to learn how to perform the activity.
051-Ability to learn firefighting procedures and techniques.
052-Ability to learn job-related rules and regulations.
065-Ability to perform arithmetic computations (for example, add, subtract, multiply, divide) to solve work problems (for example, number of hose lengths needed to reach a fire).
072-Ability to analyze mistakes to avoid repeating them (for example, to review fire scene errors after the incident has concluded).
088-Ability to recall information learned in training even when it is used infrequently.
089-Ability to recall information regarding specific events and activities (for example, tactics used in prior fire scenes).
Physical Abilities
022-Ability to use muscular force to lift, push, pull, drag, carry, objects, materials, equipment and/or people.
023-Ability to use muscular force to physically control victims as needed (for example, during roof rope rescue).
024-Ability to exert maximum muscular force to use equipment or perform other activities (for example, when using hook, axe, carrying a hose).
025-Ability to exert muscular force quickly to initiate action (for example, to start a chain saw, force a door).
026-Ability to maintain a high level of physical effort (for example, advancing hose line) under difficult environmental conditions (for example, heat, smoke, darkness).
027-Ability to bend, twist, stretch, and reach with the body, arms, and/or legs (for example, on a fire escape, entering a window).
031-Ability to coordinate the rapid movement of arms, legs, and/or the torso while the entire body is in motion (for example, climbing a ladder).
Noncognitive Characteristics
036-Willingness to adapt to and become a member of an established team.
038-Willingness to request assistance from a co-worker or supervisor when necessary to complete an assignment.
039-Willingness to offer information and/or assistance and information to co-workers when it appears necessary or when it would facilitate task accomplishment.
041-Willingness to do one's share of the work including performance of undesirable tasks.
043-Willingness to show respect toward those with more experience or in a position of authority.
045-Willingness to accept responsibility for one's own actions.
046-Willingness to comply with assignments, commitments, requirements, and/or instructions regardless of personal feelings about a situation.
047-Willingness to maintain appropriate attention to detail and persist in work activities in order to complete work in a safe, effective and timely manner.
048-Willingness to work without direct supervision.
049-Willingness to maintain high standards of ethical conduct for self and others.
050-Willingness to devote time and effort to all aspects of the job, including those that are routine in nature.
094-Willingness to risk harm to self to attempt to ensure the safety of others.
095-Willingness to obey orders promptly.
096-Willingness to accept and follow all rules and regulations.
097-Willingness to follow all safety rules, use all safety equipment, and avoid unnecessary risk.
102-Willingness to seek training or other assistance to ensure needed improvements in job performance are made.
103-Willingness to master work activities and continue learning throughout career.

[13] At a subsequent step in the project, it was determined that abilities related to (017) concentration, and (018) attention would be better addressed in the non-cognitive domain.

104-Willingness to accept constructive criticism without becoming offended.
105-Willingness to ask questions even when to do so indicates lack of knowledge or understanding.
111-Ability to convey a professional and trustworthy image; to create a positive impression.
112-Willingness to maintain appearance within department standards.
115-Ability to maintain control of personal reactions and impulses while taking charge of or handling a disagreeable or dangerous situation.
122-Ability to interact effectively with other people.
125-Ability to interact with people of both sexes and of different races/ethnicities, cultural or religious beliefs/practices, sexual orientation, and/or socioeconomic status in a fair and respectful manner.

Task Category Importance Ratings

In the final section of the Linkage survey, participants were asked to review the 17 Task Categories and the core tasks listed within each category. Then, they were instructed to distribute 100 points across the Task Categories, based on the relative importance of each category to Firefighter job performance. Appendix L shows the specific instructions. These ratings were used later in developing test specifications.

Table 4 shows the resulting means of the points allocated to each Task Category by the SMEs. The results indicate that Engine Company Operations, Victim Removal, and Search received the highest point allocations (were considered more important) relative to other Task Categories. Conversely, Clean Up/Pick Up, Inspection of Buildings/Hydrants/Alarm Boxes, and Overhaul were identified as the least important Task Categories.

Table 4. Firefighter Task Category Importance Rating Results

Task Category	Mean Relative Importance (%)	SD
1. Station & Equipment Maintenance/Chores	5.29	3.02
2. Initial Response to Incident/Driving	5.47	2.17
3. Size-Up & Initial Actions	7.29	4.78
4. Ladder Operation	7.34	6.52
5. Climbing & Portable Ladder Activities	6.14	2.87
6. Building Entry	7.15	1.97
7. Search	7.49	2.33
8. Victim Removal	7.60	2.90
9. Ventilation	6.08	2.77
10. Engine Company Operations	10.10	5.08
11. Overhaul	3.35	1.55
12. Clean Up/Pick Up	2.44	1.71
13. Inspection of Buildings/Hydrants/Alarm Boxes	2.76	1.76
14. Rescue/Extrication	6.92	2.72
15. Providing Medical Assistance	4.95	2.56
16. Training	6.13	3.16
17. House Watch Duties	3.51	1.94
N=35 raters		

Note: ratings were rescaled to total 100 for raters whose ratings did not total exactly 100.

Analysis of Reading Demands and Learning Process

The above described job analysis results indicated that reading comprehension and learning abilities are important for the successful performance of FDNY Firefighter tasks (see Table 3, abilities 001, 002, 003, 007, 008, 009, 021, 051, 052, 072, 088, and 089; also see Appendix M)

In anticipation that reading and learning abilities would potentially be included for assessment in the new test, additional information about the FDNY Firefighter job was gathered, such as the job-related documents Firefighters are required to read and description of the manner in which Firefighters acquire and learn information while in training and on the job.

Analysis of Documents Read by Firefighters

To aid in the design of the Firefighter test, particularly with respect to reading comprehension, an analysis was conducted to gauge the reading level of materials FDNY Firefighters read in academy training and on the job. The purpose of this analysis was to help ensure the new Firefighter test was calibrated to an appropriate reading level consistent with the materials encountered on the job.

Six document types representing the majority of probationary training materials Firefighters use were obtained in electronic format, including: Probationary Fighter Manual, Engine Company reading list, Ladder Company reading list, certified first responders-defibrillation (CFR-D) manual, inspection forms and building inspection forms. Further details regarding the documents included in the analysis are provided in Appendix M, Table M-1.

A reading level ("readability") analysis of the documents (a total of 4,282 pages) was conducted as a means of evaluating the relative difficulty of the materials. Two readability indices were computed for the collection of reading materials: Flesch-Kincaid[14] and SMOG.[15] These indices yield alternative estimates of reading ease/difficulty that are correlated with an approximate educational grade level. They differ in results produced because they use different parameters to predict reading level.

The analysis yielded Flesch-Kincaid values ranging from grade level 8.3 to 15.9, and SMOG index values ranging from grade 11.7 to 13.4. A full delineation of the resulting indices is provided in Appendix M, Table M-2. These findings would guide subsequent creation of test materials, ensuring that the test followed a reading level no greater than the range of the materials firefighters actually encounter in training and on the job.

---

[14] The Flesch-Kincaid index estimates reading grade level (G) for an average student based on the number of words, syllables and sentences in the document, as follows: $G = (11.8*(B/W)) + (.39* (W/S)) - 15.59$
Where: G = Grade Level; W = Number of words per document; B = Number of syllables per document; and S = Number of sentences per document.

[15] The SMOG index estimates reading grade level (G) based on polysyllable words to estimate reading level at 100% comprehension, as follows: $G = 1.0430 * \sqrt{C} + 3.1291$
Where: G = Grade Level; and C = Number of complex words (3+ syllables) per document

Firefighter and Trainer Interviews about the Learning Process

Additional information was collected to aid in design of the test with respect to the ability to Learn and Apply Information. In doing so, information was also gathered regarding the use of reading materials and how basic arithmetic may be used on the job. To this end, a structured interview format was developed and a series of telephone interviews were conducted with 10 FDNY Firefighters and 5 academy instructors to gather information regarding the ways in which Firefighters acquire information and learn the job, both during initial academy training, and thereafter. Testing experts from PSI, DOJ and the Vulcan Society each conducted interviews with Firefighters representing engine and ladder companies from a cross-section of boroughs, as well as instructors, all representing different racial/ethnic groups. Table 5 summarizes the characteristics of the interview sample with respect to job title, role, borough, unit type, race/ethnicity, and gender.

Table 5. Description of Participants in the Learning Process Interviews

		Frequency	Percent
Job Title	Firefighters	13	86.7%
	Lieutenants	2	13.3%
Role	Non-Instructor	10	66.7%
	Instructor	5	33.3%
Borough	Brooklyn	4	26.7%
	Bronx	5	33.3%
	Manhattan	2	13.3%
	Queens	2	13.3%
	Staten Island	2	13.3%
Unit Type	Engine	6	40.0%
	Ladder	7	46.7%
	Battalion	2	13.3%
Race/Ethnicity	Black	6	40.0%
	Hispanic	5	33.3%
	White	4	26.7%
Gender	Female	3	20.0%
	Male	12	80.0%
Total		15	100.0%

Appendix N contains the structured interview forms used for the Firefighters and trainers, respectively. Key findings from the interviews indicated that with regard to Reading and Learning: (a) all Firefighters interviewed spent time reading the probationary training manual and reading on the job; (b) most respondents indicated that reading information contained in manuals and the reading list was also linked (occurred together) with other learning sources (lectures, demonstrations, videos); (c) nearly all Firefighters interviewed indicated that it would be difficult or impossible to pass the academy ("Probie School") without reading or referring to the training manual; and (d) reading is an important mode of learning various key parts of the job, consistent with the job analysis survey results.

With regard to the use of basic arithmetic: (a) all interview respondents indicated Firefighters must rely on basic arithmetic skills without the use of a calculator; (b) the level of arithmetic required is basic and test items should reflect this finding; i.e., test items should address basic arithmetic problems without the use of a calculator; and (c) a limited number of examples of arithmetic were provided (i.e., ladder heights, hose lengths and pressure).

Development of Job Analysis-Based Performance Measures

The job analysis results served as a basis for developing measures of Firefighter job performance to serve as criteria for evaluating the predictive effectiveness of the new Firefighter CBT. These included (1) job performance ratings specifically developed for this project on the basis of the 2011 FDNY Firefighter job analysis; and (2) Probationary Academy performance data.[16]

Job Performance Rating Booklet

Further expanding the job analysis, a research-only Job Performance Rating Booklet was developed specifically for use in the validation study as a means for Lieutenants and Captains to evaluate the job performance of the Firefighter incumbents who were selected to take the CBT. The rating instrument was developed by PSI as part of the job analysis process, in consultation with the parties' testing experts, following several steps, including: development of job-analysis based performance rating dimensions and scales; Firefighter/Officer focus group review; pilot administration; and revision and assembly of final rating booklets.

The rating instrument was designed to assess incumbent Firefighters' job performance with respect to important work areas that represent important tasks, abilities and work behaviors identified in the 2011 job analysis of the FDNY Firefighter position. The initial draft instrument contained 18 work area dimensions and several summary dimensions reflecting work habits and outcomes, as well as overall performance; all based upon the job analysis findings. In addition, background information questions, rating scales and instructions were drafted.

On May 20, 2011, PSI conducted a focus group session with ten FDNY Fire Lieutenants and Captains at the FDNY Academy. The purpose of the half-day session was to review the job performance rating dimensions, rating scales, and instructions, and to confirm that they were clear, accurate, complete, useable, and related to the job. As a result of the meeting, several edits were made to the background questions, rating dimensions and instructions.

The rating dimensions, scales and instructions were assembled into a final rating booklet comprised of five sections, including:

> Section 1 – Background Information: The evaluator was asked to provide general descriptive information regarding his/her work assignment and experience, time supervising the Firefighter/ratee, and overall familiarity with the Firefighter's performance.

[16] Employee records of absenteeism and medical leave maintained by the FDNY were also examined as potential job performance criterion measures. However, these did not prove to be useful for the study because there was little or no variation among Firefighters on these data.

Section 2 – Work Area Ratings: The evaluator used a 6-point rating scale (ranging from 1=Poor to 6=Excellent) to describe the Firefighter's effectiveness in each of 18 Work Areas. The rating scale that was used to evaluate performance in each work area is shown in Figure 1a.

Section 3 – Work Habits and Outcomes: The evaluator used a 4- or 5-point rating scale, depending upon the specific rating area, to describe any complaints regarding the Firefighter's performance, attendance and tardiness.

Section 4 – Overall Job Performance: The evaluator used the same 6-point effectiveness rating scale to describe the Firefighter's overall job performance as was used for the Work Area Ratings (see Figure 1a).

Section 5 – Relative Ranking: The evaluator used a 6-point scale to rank the Firefighter's performance compared to that of other Firefighters in the unit with respect to each of 4 summary dimensions and overall performance rating (see Figure 1b).

The job performance rating dimensions, linked directly to the results of the job analysis, included in sections 1-5 of the rating booklet are listed in Table 6. The final Job Performance Rating Booklet is shown in Appendix O.

Table 6. Job Performance Rating Dimensions

Work Areas Rated on 6-point Effectiveness Scale	
1.	**Size-up and Initial Response** How effectively the Firefighter: evaluates the fire or incident scene to determine what actions initially should be taken; communicates and takes appropriate initial action based on prior preparation as well as size-up.
2.	**Building Entry, Ladder Usage, Ventilation, & Overhaul** How effectively the Firefighter: forces doors or otherwise enters buildings to search for and rescue victims and to provide access to the fire for offensive firefighting; stabilizes ladder trucks operating aerial and tower ladders to rescue victims; raises, sets up, and climbs ladders; provides access for ventilation; operates master stream devices, etc.; opens or breaks open windows; chops or cuts bolts in roofs; breaks through walls or doors; hangs fan in windows or doors to remove heat, smoke and gas from burning buildings; opens up walls and ceilings; cuts or pulls up floors; moves or turns over debris to check for hidden fire that could rekindle or spread using hooks, axes, and saws.
3.	**Using Pumps, Hydrants and Hose Lines** How effectively the Firefighter: connects and hooks up engine to fire hydrant; operates pumps to supply water of appropriate pressure and volume for firefighting; uses tools (e.g., wrenches, couplings) necessary for proper setup and operation of hose lines and hydrants; stretches and operates hose lines.
4.	**Search and Rescue/Extrication** How effectively the Firefighter: searches fire or assigned area to locate victims and obtain further information about the fire, following standard search procedures; extricates victims from vehicles, cave-ins, collapsed buildings, subways, elevators or other entrapments to save lives using shovels, torches, drills, saws, jacks, hurst tools, air bags, and other equipment.
5.	**Taking Up & Salvage** How effectively the Firefighter: picks up and returns equipment to apparatus so that the company can go back in service; takes into account nature of property and attempts to minimize damage as appropriate when performing clean-up; moves and covers property; covers holes in buildings; redirects or cleans up water in order to minimize damage; uses covers, ropes, staple guns, and other tools.

6. **Providing Victim and Medical Assistance**
How effectively the Firefighter: provides assistance to victims; provides first aid and direct medical assistance to persons requiring emergency attention.

7. **Inspection**
How effectively the Firefighter: inspects buildings for code violations or hazards on a periodic basis or during the course of activities; inspects hydrants for operational use.

8. **House Watch**
How effectively the Firefighter: stands watch to receive incoming alarms and information; answers phones; monitors access to the station house.

9. **Station and Equipment Maintenance**
How effectively the Firefighter: inspects, cleans, and maintains apparatus, equipment carried on the apparatus, and personal gear and equipment; performs routine housekeeping chores and "committee work".

10. **Professional Development**
How effectively the Firefighter: devotes the time and effort necessary to expand job skills and capabilities.

11. **Focus and Persistence**
How effectively the Firefighter: persists in activities; takes appropriate initiative; attends to details; puts forth extra effort as appropriate to complete job tasks.

12. **Teamwork**
How effectively the Firefighter: helps and cooperates with other team members.

13. **Interpersonal Effectiveness**
How effectively the Firefighter: interacts with other people including the general public.

14. **Compliance with Rules and Procedures**
How effectively the Firefighter: follows organizational rules, procedures, and standards of conduct.

15. **Physical Performance**
How effectively the Firefighter: performs work activities that require physical strength, endurance, and flexibility.

16. **Learning and Applying Information**
How effectively the Firefighter: applies what he/she has learned from a variety of sources to perform the job.

17. **Communication**
How effectively the Firefighter: listens, understands, and conveys the information needed to accomplish job tasks, orally and by the use of brief written notes.

18. **Problem Solving**
How effectively the Firefighter: solves work problems, makes decision, and takes appropriate action.

- **Overall Effectiveness:** Considering the Firefighter's overall job performance, how would you describe the overall effectiveness of the Firefighter during the past 12 months?

Summary Dimensions Rated on 6-point Ranking Scale

1. **Using Technical Firefighting Capabilities** – Safely and efficiently performs on-scene technical firefighting and emergency operations; uses experience and up-to-date technical knowledge and skill to accomplish firefighting duties and tasks effectively.

2. **Getting Along with Others** – Is a good teammate; interacts effectively with the public; works well with co-workers by coordinating and cooperating; treats others with respect and acts in a considerate manner.

3. **Dependable Organizational Citizen** – Acts professionally and responsibly; is reliable and someone you can count on; willingly assists others and does fair share; focuses on goals of the department.

4. **Stamina and Strength** – Uses physical fitness, agility, and prowess to perform tasks safely and effectively.

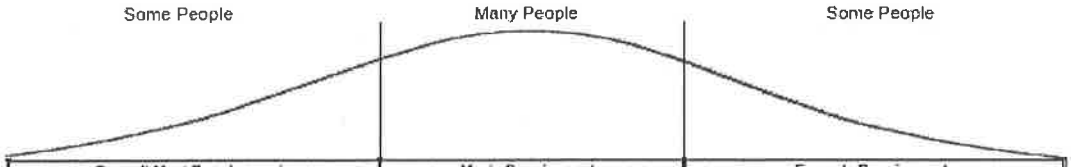
- **Overall Ranking** Compared to all Firefighters in your company who perform similar work activities, how would you rank this Firefighter's overall job performance during the past 12 months?

Work Habits
1. How many complaints are you aware of about the Firefighter's performance during the **past 12 months?** (For example, by co-worker, supervisor, resident, other agency.)

Figure 1. Job Performance Rating Scales

a. Effectiveness Scale

Use the following rating scale to describe the effectiveness of each Firefighter in each of 18 work areas during the past 12 months.



Doesn't Meet Requirements		Meets Requirements		Exceeds Requirements	
1 *Poor*	2 *Needs Improvement*	3 *Just Adequate*	4 *Fully Acceptable*	5 *Very Good*	6 *Excellent*
• Often has trouble performing effectively in this area. • Fails to perform acceptably in this area; produces unsatisfactory results. • Work is incomplete, inaccurate, or insufficient; makes frequent or serious errors. • Work in this area creates additional unnecessary work or problems. • Does not perform work in a timely manner in this area.		• Consistently performs acceptably in this area. • Achieves satisfactory results in this area. • Work is sufficiently accurate and complete in this area. • Meets basic requirements for work speed and efficiency in this area.		• Consistently performs work at a level that exceeds basic requirements in this area. • Achieves superior outcomes in this area. • Rarely or never makes mistakes in this area. • Assists and corrects others' mistakes in this area. • Performs work quickly and without wasted effort in this area. • Serves as a role model to others for work in this area.	

NA: Not Able to Observe -- insufficient opportunity to observe person during the past 12 months *(circle NA).*
NP: Not Part of Job – not a part of this person's job *(circle NP).*

Note: Use the same standards to rate newer Firefighters as longer-term Firefighters. If the Firefighter's performance is less than acceptable for any part of a specific work area, your rating should reflect this. If you have been unable to observe the Firefighter performing one of the listed work areas, do <u>not</u> describe the individual's performance. Instead, circle "NA" (Not Able to Observe) for that particular item. Similarly, if the work area is not part of the Firefighter's job, circle "NP" (Not Part of Job) for that particular item.

b. Relative Ranking

Compared to all Firefighters in your company who perform similar work activities, how would you rank this Firefighter's job performance in the work area of <_____> during the past 12 months? In other words, if you were to rank the job performance of all the Firefighters in your company from best to worst in the work area of <_____>, where would this person fall?

1 = Bottom 10% of Firefighters in your company
2 = Bottom 30% of Firefighters in your company
3 = Bottom 50% of Firefighters in your company
4 = Top 50% of Firefighters in your company
5 = Top 30% of Firefighters in your company
6 = Top 10% of Firefighters in your company

Probationary Firefighter Training Academy Performance

Performance in the probationary Firefighter training academy served as an additional measure of initial performance as a new Firefighter. During a 23-week period, newly hired Firefighters attend training academy lectures and demonstrations, and learn the material in the Probationary Firefighter Training manual in order to perform the job. During the academy training, Firefighters must demonstrate their proficiency with the training curriculum in a series of examinations, including (1) written quizzes, (2) practical exercises in which Firefighter trainees demonstrate their knowledge and ability and are rated by an instructor; (3) a written mid-term exam, and (4) a final exam. Firefighters' scores on these four types of academy performance measures were collected to serve as potential criterion measures against which to evaluate the criterion-related validity of the new Firefighter test.

Summary

A comprehensive job analysis was conducted for the entry-level FDNY Firefighter job. The study incorporated the input of over 400 subject matter experts comprised of carefully drawn samples of Firefighter incumbents and officers representing the five boroughs, engine and ladder companies, and racial/ethnic and gender groups.

Consistent with prior studies of the Firefighter job and industry best practices, formal quantitative ratings of job tasks, abilities, and characteristics were obtained and analyzed to identify the core tasks performed and the abilities and characteristics required to learn and successfully perform the job. Analyses of these ratings indicated the results were reliable and consistent with prior studies of the FDNY Firefighter job. The study resulted in the identification of 108 core job tasks performed, and 49 abilities and characteristics that are required Day-1 for job success and linked to core tasks. The study also yielded information regarding the range of reading level for documents that Firefighters use, as well as the process for acquiring new information and learning in the academy and on the job.

The job analysis results served as a basis for developing a Firefighter performance evaluation tool for use as a criterion measure in a subsequent validation study. Training academy data were also identified to serve as test validation criterion measures.

Furthermore, the job analysis results (abilities and characteristics) provided an important foundation for designing test specifications and developing a new Firefighter selection testing program for FDNY.

CHAPTER 3: DEVELOPMENT OF A NEW FIREFIGHTER TEST

Introduction[17]

This chapter summarizes the design and development of a new test for entry-level Firefighter selection at the FDNY. The test was designed by PSI specifically to measure cognitive abilities and other characteristics identified in the FDNY Firefighter job analysis as essential for successful performance of Firefighter job tasks, and required prior to entry into the probationary Firefighter Academy. Several equivalent alternate forms of the test were developed to maintain the security of the test by limiting and scrambling the exposure of the test content during administration to candidates.

The new test incorporated advanced technology features to provide a realistic, objective, and secure assessment. A computer-based test (CBT) system was utilized to deliver and score the test in a standardized and reliable manner, with enhanced security features such as item randomization. Many of the items incorporated multimedia (audio/video) formats to simulate the manner in which Firefighters learn and apply new information; traditional multiple-choice written questions were also included.

The following sections describe the test design and development procedure, which took place between February and November 2011. A summary of the final version of the Firefighter CBT is also presented.

The test development methodology was designed to be consistent with professional standards and principles (AERA, APA, NCME, 1999, *Standards for Educational and Psychological Testing*; SIOP, 2003, *Principles for the Validation and Use of Personnel Selection Procedures*; EEOC et al, 1978, *Uniform Guidelines on Employee Selection Procedures)*.

Test Design Founded on Job Analysis

Design specifications for the new test were developed on the basis of the job analysis results. Attention also was given to the broader psychometric literature on the measurement of abilities and characteristics, along with review of other selection tools used in screening Firefighter candidates. This work resulted in producing a test specification that was used to define the content areas that would be measured on the new test and the development of test items.

Assessment Dimensions

The first step in designing the test was to identify assessment dimensions to serve as the content areas for which test questions would be developed. Assessment dimensions were developed directly from the job analysis results by grouping together the "core" abilities and characteristics identified in the FDNY Firefighter job analysis. The assessment dimensions were developed by PSI, with assistance from DOJ and Vulcan Society testing experts in meetings conducted in February and March 2011. As a result of these meetings and subsequent test development and review sessions, the abilities and characteristics identified in the job analysis were grouped into

[17] Uniform Guidelines Section 14 C.(3); 15 B.(7); C.(4); D.(6).

ten major content areas. The ten areas also were linked to literature-based constructs falling within cognitive[18,19] and noncognitive[20] domains. The final assessment dimensions included:

Cognitive Abilities:

- Reading Comprehension
- Ability to Learn and Apply Information
- Basic Arithmetic
- Listening Comprehension
- Reasoning

Noncognitive Characteristics:

- Conscientiousness
- Agreeableness
- Emotional Stability
- Interpersonal Competence
- Honesty/Socialization

The new test was not designed to address physical abilities documented during the job analysis process. These requirements of the FDNY Firefighter job will be assessed by the City in a separate testing instrument.[21]

Test Specifications

Once the assessment dimensions were identified, specifications for test item development were developed by PSI, in consultation with the DOJ and Vulcan Society testing experts. In doing so, consideration was given to utilizing test formats and item types that were:

- Realistic and similar to how individuals learn and perform the job, incorporating multiple modes of information (lecture, demonstration, observation, reading);
- Multimedia-based and not overly reliant upon reading comprehension, particularly for assessment dimensions where reading was not the specifically targeted ability;
- Comparable in reading level to documents used by FDNY Firefighters (i.e., for Reading Comprehension test passages);
- Reliable, objective, practical and amenable to automated delivery and scoring with a candidate pool numbering in the tens of thousands; and
- Resistant to cheating and harvesting (stealing) of content.

The resulting test specifications for the cognitive assessment dimensions took the form of a video-based simulation, depicting an entry-level Firefighter's initial assignment to the FDNY

---

[18] Ackerman, P.L., & Heggestad, E.D.(1997). *Intelligence, Personality, and Interests: Evidence for Overlapping Traits*. Psychological Bulletin. Vol. 121, No. 2, 219-245.

[19] Fleishman, E.A., Quaintance, M.K., & Broedling, L.A. (1984). *Taxonomies of Human Performance: The description of human tasks*. Orlando, FL: Academic Press, Inc.

[20] Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. ln N. R. Anderson, D . S. Ones, H . K. Sinangil, & C. Viswesvaran (Eds.), Handbook of 'Industrial, Work & Organizational Psychology: Vol. I (pp.233-277). London and New York: Sage.

[21] The Candidate Physical Ability Test (CPAT) is presently used by the City to assess Firefighter candidate physical capabilities.

training academy and attendance at training sessions, supplemented by test items simulating excerpts from an operations manual focusing on reading material related to the simulated lesson. Specifications representing the noncognitive assessment dimensions included traditional multiple-choice questions of several types (biographical, work attitude, and workplace personality); 14 noncognitive facets were identified to represent the noncognitive assessment dimensions. Computer-based test (CBT) delivery of test questions was specified to support controlled delivery of multimedia question types that incorporate video and graphic images. CBT also enables secure delivery of alternate test forms and scrambling of test questions, which helps to foil cheating and content harvesting efforts.

Development of Multimedia Test Items and Materials

On the basis of the test specifications and other information (e.g., samples of FDNY job and training materials) PSI drafted scenario scripts and test items referencing fictitious equipment. A training simulation was created to provide a unified theme within which to address the cognitive assessment dimensions. Noncognitive items were written using established formats and question themes to assess literature-based constructs linked to the various core Firefighter characteristics within this domain.

Training Simulation Design

A training simulation was designed by PSI to assess the cognitive assessment dimensions by creating a fictional scenario with novel content pertaining to fictitious equipment and procedures. Care was taken to ensure that there would be no prior knowledge of equipment, procedures or operating principles needed in order to learn and apply the information from the training simulation. In creating content for the scenario, the FDNY Firefighter Probationary Academy Training Manual was reviewed to identify examples of equipment, materials, technical information, and operating procedures and guidelines that could be modeled in a novel way. To this end, scenarios were created for two training lessons: (1) Operation, technical properties and use of a fictitious piece of equipment, and (2) Technical properties, detection, and communication procedures for a fictitious hazardous chemical.

Once two scenarios were created, scripts were drafted for actors to play an instructor and student for purposes of filming an enactment of the scenarios and serve as stimulus materials for the test. The fictitious equipment and various props used in the scenarios were also constructed by PSI staff. Test items were written and reviewed by Ph.D. and Master's level staff with degrees in Industrial Psychology and extensive experience in test development. Test questions were written for each of the assessment dimensions, incorporating the range of item types included in the test specifications. These included multiple-choice single answer; multiple-choice multiple-answer; graphical item stems; graphical item response options; hot spot (click on a picture) items; and drag-and-drop items that required respondents to click-on and move things to indicate the appropriate order. Reading comprehension passages were written to simulate an operations manual for the fictional equipment and adjustments to sentence length and use of polysyllable words were made to ensure the reading level was in the range of FDNY Firefighter documents (reading level ranged from grade 9.9 to 11.5 for the reading comprehension passages included in the final CBT Form A). The simulated lectures and demonstration were filmed on June 1, 2011, and post-production of the media files took place over the following two weeks. A Training

Guide was developed for use by examinees during the CBT to briefly study before each video lesson, to take notes during the video lesson, and to refer to in answering the test questions.

The draft scenario scripts, test items and videos were posted on a secure web server to provide the DOJ, Vulcan Society and Special Master testing experts an opportunity to review and comment on the materials prior to a formal review session in June 2011.

An overview of the final version of the Training Simulation Test-Form A is shown in the last section of this chapter.

Noncognitive Items

Noncognitive items were developed using a structured process that involved the following steps.

- PSI conducted a literature review to identify promising item types and formats for use. Multiple item types were identified and used (e.g., subtle vs. contextualized items; biographical vs. self-description).
- PSI prepared guidelines for item writers, describing effective approaches for writing attitudinal, biographical, and self-description items.
- Item writers (Industrial/Organizational Psychologists) drafted items covering the targeted assessment dimensions and associated constructs.
- PSI project team members selected items from the resulting item pool that represented a balanced array of item types, assessment dimension coverage, and related noncognitive construct coverage.
- DOJ and Vulcan Society testing consultants reviewed and provided feedback for revision of the selected draft items.

Fourteen (14) noncognitive scales were developed to represent facets of the five noncognitive assessment dimensions. An overview of the noncognitive portion of the final validated test (Noncognitive Test Form A) is shown in the last section of this chapter.

PSI staff authored the test items, media files and examinee instructions into PSI's proprietary CBT system. The content was posted and subjected to PSI quality assurance reviews in preparation for administration of a pretest of the new materials in July 2011.

Testing Expert Review

A formal review of the draft videos and test items was conducted with testing experts representing DCAS/PSI, DOJ, the Vulcan Society and the Special Master in Chicago, on June 21-22, 2011. The objective of the session was for the experts to review and confirm appropriateness of test instructions, scripts and video test prompts and to identify and suggest any needed edits by consensus. PSI prepared an "item book" for each of the session participants, which included a complete copy of all test questions and related exhibits. PSI presented the videos to the participants using a projector and screen.

The experts recommended edits and enhancements, such as adding pauses and graphical cues in the videos; modifying and clarifying some of the instructions; clarifying certain images and

exhibits associated with test questions; and miscellaneous edits to certain test questions. In addition, the experts agreed that additional biographical questions should be written.

PSI executed the recommended modifications and made additional refinements to the videos, instructions and questions. The videos were edited and narration for the test instructions was re-recorded. The updated items and videos were posted to a secure web site to provide the testing experts an opportunity to review the revised content.

Description of the Firefighter CBT

An experimental version of the Firefighter CBT was assembled for pilot testing and use in validation phases of the project. This experimental version contained extra test items to allow for the subsequent selection of test item sets that best predict job performance, while minimizing subgroup differences.

The final version of the Firefighter CBT is summarized in Table 7. The information presented reflects the final validated version (Form A) of the test. The cognitive portion of the test contained 57 items; the noncognitive portion contained 65 items. The total examination time was projected at four hours.

Table 7. Overview of the Firefighter CBT

Cognitive (57 items)	**Protocol**	**Assessment Dimension**
Video Lesson *(26 items)* Part 1 Lecture – Equipment Operation and Use	*Examinee is provided a printed outline of the training material to review; untimed.* *Then examinee watches video lesson, taking notes on the training outline; 2 replays allowed; then answers several questions.*	• Ability to learn and apply information • Listening comprehension • Reasoning
Part 2 – Narrated slide show on Equipment safety	Examinee watches narrated slide show, taking notes on the training outline; then answers several questions.	
Part 3 – Student-Instructor Questions & Answers	Examinee watches three, 1 and 1/2 minute video clips, taking notes on the training outline; 2 replays allowed; answers several questions.	
Operations Manual *(31 items)*	Examine reads a 1-page excerpt from an operations manual for the equipment, shown on the right side of the computer screen; then answers several questions regarding the passage; a total of 3 different passages are presented; some content overlaps with the lecture.	• Reading Comprehension • Basic Arithmetic • Reasoning
Noncognitive (65 items)	**Protocol**	**Assessment Dimension**
Background Questions *- Dependability* *- Activity* *- Agreeableness* *- Even Tempered* *- Low Anxiety* *- Self Esteem*	Examinee is presented a series of multiple-choice questions (biographical, behavioral description, endorsement)	• Conscientiousness • Agreeableness • Emotional Stability • Interpersonal Competence • Honesty/Socialization

Summary

A new computer-based test for entry-level Firefighters was designed to assess abilities and characteristics that were identified in the 2011 job analysis as important or critical to successfully learn and perform the FDNY Firefighter job. The new test incorporated multimedia test item formats in a CBT system to enable a realistic simulation of learning and applying information to solve problems in a manner similar to what might occur in training and on the job. The new test also included noncognitive items to assess candidate background characteristics and work approach. The CBT system provided enhanced security in the delivery of alternate test forms and scrambled content.

A draft form of the test (Form A) was produced for pretesting and validation. The draft form contained additional content to allow for the possibility of selecting portions of the test that are equally valid and which have less adverse impact for racial/ethnic and gender minority group candidates.

CHAPTER 4: DOCUMENTATION OF VALIDITY EVIDENCE

Introduction[22]

This chapter describes the validation strategies and studies conducted to establish evidence of validity for the new Firefighter CBT. The validation studies were designed to follow established professional standards and principles for test validation promulgated in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) and the *Principles for the Validation and Use of Personnel Selection Tests* (SIOP, 2003) and to comply with the *Uniform Guidelines*.

Three strategies were followed to establish validity evidence for the new Firefighter CBT, which are specified in the *Uniform Guidelines*: (1) content validity (sec. 14C); (2) criterion-related validity (Sec. 14B); and (3) construct validity (Sec. 14D). Evidence of content validity was established for the cognitive portion of the CBT by collecting data showing that the content of the test represents important aspects of performance on the FDNY Firefighter job. Criterion-related validity evidence was established by gathering empirical data demonstrating that CBT scores are significantly correlated with important aspects of FDNY Firefighter job performance. Construct validity evidence was established by collecting data showing that scores on the CBT designed to measure abilities and characteristics that are important for successful performance in the FDNY Firefighter job are significantly correlated with other established tests of the same or similar abilities and characteristics.

An additional empirical study was conducted to construct alternate forms of the CBT and to document their equivalence to the original Firefighter CBT form developed and validated in the above studies.

The methodology and results of each validation strategy are described in the following sections of this chapter.

[22] Uniform Guidelines Section 14 B. (1),(3),(4),(5),(8); C.(1),(4),(5),(6),(7),(8),(9); D.(1),(3); Section 15 B.(1),(2),(5),(6),(8),(9),(13); C. (1),(2),(5),(6),(9); D. (1),(2),(3),(7),(8),(10).

Content Validity Study

Content validity evidence for the Firefighter CBT was compiled in a study that was designed to establish a linkage between the content of the training simulation (cognitive) portion of the test and important elements of the FDNY Firefighter job. The noncognitive (background questions) portion of the CBT was not included in the content validation study. During the study, the CBT was administered to the participating Firefighters and Officers to ensure they were familiar with the test, and also to pilot the CBT.

The study entailed a series of steps, including selection of a sample of FDNY Firefighters and Officers to serve as job experts; development of a content validation rating form; conducting a content validation workshop and testing session; and analysis of content validation ratings. The Firefighter's and Officer's responses to the CBT items were also reviewed and served as a basis for making refinements to instructions and testing time.

Date(s), Time(s) and Location(s): A content validation and test review session was conducted on July 22, 2011 at the DCAS testing center located at 210 Joraleman Street, Brooklyn, NY.

Sampling Plan

PSI selected 30 participants from an electronic file listing all Firefighters and officers employed by FDNY. The sample was designed to include:

- 15 Firefighters with 15 or more years of tenure at FDNY (the 15 years of tenure qualification was utilized to ensure that the sample would avoid all Firefighters selected to participate in a subsequent criterion-related validation study of the new Firefighter test);
- 15 fire captains and lieutenants;
- Engine and Ladder companies;
- 5 boroughs; and
- Racial/ethnic groups.

Characteristics of the session participants are summarized in Table 8 with respect to job title, race, borough, unit, and experience. While the sampling plan sought a varied group of participants, it was not designed to yield a representative sample of the FDNY. The sample was intentionally deigned to avoid duplication of the large number of firefighters selected to participate in a subsequent criterion-related validity study. For example, all female FDNY Firefighters were selected to participate in a subsequent criterion-related validity study and thus none were available to participate in the content validation session.

Table 8. Characteristics of the Content Validation Session Participant Sample

	Frequency	Percent
Job Title		17.2
Captain	5	
Firefighter	14	48.3
Lieutenant	10	34.5
Race		
Black	9	31.0
Hispanic	10	34.5
White	10	34.5
Borough		
Bronx	5	17.2
Brooklyn	8	27.6
Manhattan	5	17.2
Queens	7	24.1
Staten Island	4	13.8
Unit		
Battalion	1	3.4
Engine	17	58.6
Ladder	9	31
Other	2	6.8
Total	29	100

Experience	N	Minimum	Maximum	Mean	Std. Deviation
Years served	29	9	32	19.62	5.691
Years served in current title	29	3	33	13.72	7.745

Content Validity Rating Form

PSI drafted a rating form to capture job experts' (i.e., the above sample of Firefighters and officers) judgments with respect to the job-relatedness of the Training Simulation (cognitive) portion of the new Firefighter test. Specifically, the rating form was designed to elicit data to determine whether the CBT represents the type of content encountered on the job and allows the test taker the opportunity to demonstrate abilities needed on the job; i.e., showing that the content of the CBT is representative of important aspects of performance on the Firefighter job. This approach was based on an established principles and practice (SIOP, 2003; Goldstein, Zedeck & Schneider, 1993).

As noted in Chapter 2, the 2011 job analysis of the FDNY Firefighter job served as a basis for defining important aspects of the job. The job analysis identified 18 core cognitive Firefighter abilities as very important or critical for successful performance of Firefighter job tasks. The content validation rating form listed these 18 core abilities and asked each Firefighter and officer

to rate the extent to which Training Simulation portions of the new Firefighter test required each listed core ability, using the following rating scale:

> *To what extent is this ability required to perform the test exercise?*
>
> **2 = Critical:** this ability is essential to perform successfully on the test exercise. Successful performance on the exercise would be extremely difficult or impossible without this ability.
>
> **1= Important:** this ability is very helpful to perform successfully on the test exercise. Successful performance on the exercise would be difficult without this ability.
>
> **0 = Not Important:** this ability is not needed to perform successfully on the test exercise. Successful performance on the exercise is possible without this ability.

Appendix P lists the 18 abilities included in the rating form. The form shows the CBT was divided into 4 exercises for the purposes of the content validation process. A test exercise was defined as the process and tasks that the examinee must perform to answer a set of questions (e.g., review documents, watch a video lecture, take notes, and answer questions; read pages from a manual and answer questions; and look up items in a reference table).

Content Validation Workshop

Prior to the session, the Special Master and her expert, as well as the parties' testing experts, and legal counsel met with leaders of the Firefighter and Fire Officer unions and their counsel to discuss the project and the participation of the unions' membership. Participants were informed regarding the nature and purpose of the session via a brief memo that was distributed in advance of the session; see Appendix Q.

The content validation session was convened on July 22, 2012 at the DCAS Brooklyn facility. A total of 29 Firefighters and officers attended the session (see Results section). PSI conducted a participant briefing which included an overview of the project, a description of the role of the participants, review of the session agenda, and a brief tutorial on how to use the computer testing system.

Following the briefing, participants were directed to the testing room to complete the Training Simulation portion of the draft Firefighter CBT (Form A), which contained extra items to allow for flexibility in assembling the final test form. Participants were assigned to a computer station and provided with a "Training Notebook" and an instruction sheet on how to start the CBT. In addition, five testing experts signed in and completed the test, including three from DOJ, one representing the Vulcan Society, and one representing the Special Master.

The testing session was proctored by PSI and DCAS staff. Participants reviewed a tutorial on the CBT system and completed several practice questions before completing the CBT. Participants worked at an individual pace and were instructed to return to the classroom for a break when finished.

Next, PSI facilitated a debriefing discussion with the Firefighters and Officers to obtain feedback on the features of the computer testing system. During this discussion, PSI noted a number of

system features to be considered for enhancement prior to subsequent steps in the project. These included features related to the control of the videos, prompting by the system, use of the training notebook, and other features.

Then, PSI staff facilitated a content validation rating process. During this process, PSI projected instructions onto a screen to guide participants through the completion of the content validation rating form referenced earlier. The rating process entailed the following steps:

a. Review of general instructions.
b. Overview of the rating form and scale.
c. Review of first test exercise (training notebook, videos, and set of questions).
d. Participants' independent rating of the importance of each of 18 abilities for the test exercise, recording their ratings on individual rating forms.
e. Group discussion of the CBT exercise, identifying any potential suggestions for refinement.
f. Proceeding to the next exercise and repeating steps c – e for each remaining exercise.

Following the content validation rating process, the participants returned to the testing room to complete the noncognitive portion of the CBT, which was not a subject of the content validation rating process. Participants worked at an individual pace and were dismissed upon completion of the test.

Content Validity Results

The content validation ratings completed during the session by the Firefighters and Officers were key-entered and verified, and a database was constructed for analysis. The ratings were subjected to various checks for reasonableness, including assessing the frequency of missing ratings and reviewing within-rater variance across items. No surveys were found to be problematic and, thus, all were retained for the analysis.

Summary statistics for the content validation ratings are presented in Table 9. All the test exercises were rated as "important" or "critical" (linked by at least 2/3 of raters) to a minimum of 14 abilities, and as many as 18 of the 18 core Firefighter abilities. Moreover, all 18 of the core Firefighter abilities were linked to at least one test exercise. These results support the content validity of the Training Simulation portion of the new Firefighter CBT.

The content validation session also provided an opportunity to pilot the CBT prior to the large-scale administration that would take place for the subsequent criterion-related validation study, which would entail administering the test to a large number of Firefighters. The pilot data were reviewed to gain a preliminary understanding of how candidates responded to the items, whether they had sufficient time, and whether the instructions were clear. On the basis of these data and comments received during the post-test debriefing, a number of refinements were made to instructions, test time, and points in the narrated videos.

Table 9. Content Validation Results

Core Firefighter Ability	Percent of Firefighters and Officers Rating Ability as Important or Critical for each Exercise			
	Exercise A*	Exercise B*	Exercise C	Exercise D
001-Ability to read and interpret short messages written in English (for example, notes, log entries, teleprinter tickets).	96.6	96.6	93.1	96.6
002-Ability to read routine documents written in English (for example, bulletins, articles, notices, announcements) to keep apprised of current job-related information.	96.6	100	89.7	96.6
003-Ability to read and interpret technical materials written in English (for example, instructional manuals, operating manuals, and other official FDNY documents) to learn new information and/or update job knowledge.	96.6	96.6	89.7	96.6
005-Ability to write brief notes/statements in English (for example, fill in forms, log entries, take messages) legibly, completely, and accurately.	96.6	93.1	82.8	96.6
007-Ability to listen to, and understand information on how to perform a task or series of tasks from a trainer or others (for example, instructions from a commanding officer, training information on specific steps to follow in different situations).	96.6	69.0	79.3	96.6
008-Ability to understand information presented orally in English, both in person (for example, in a training session) and from a variety of communications devices (for example, radio, phone, intercom).	96.6	79.3	75.9	93.1
009-Ability to listen to and understand people in emergency situations (for example, people who are upset, frightened, confused).	62.1	55.2	62.1	69.0
010-Ability to state ideas clearly and concisely when speaking in English (for example, giving instructions, explaining procedures, providing technical information).	75.9	72.4	72.4	75.9
016-Ability to quickly and accurately compare letters, numbers, information and objects (for example, addresses, names, radio codes) to determine if they are the same or different.	89.7	86.2	96.6	86.2
017-Ability to concentrate on the work to be performed in spite of distractions, keeping aware of one's surroundings (for example, during a highway accident).	75.9	65.5	82.8	72.4
018-Ability to remain attentive while performing routine or repetitive tasks (for example, taking up hose line).	79.3	86.2	86.2	79.3
021-Ability to observe another person performing/ demonstrating an activity to learn how to perform the activity.	82.8	69.0	65.5	86.2
051-Ability to learn firefighting procedures and techniques.	75.9	86.2	62.1	75.9
052-Ability to learn job-related rules and regulations.	89.7	93.1	65.5	79.3
065-Ability to perform arithmetic computations (for example, add, subtract, multiply, divide) to solve work problems (for example, number of hose lengths needed to reach a fire).	86.2	96.6	75.9	79.3
072-Ability to analyze mistakes to avoid repeating them (for example, to review fire scene errors after the incident has concluded).	79.3	82.8	75.9	82.8
088-Ability to recall information learned in training even when it is used infrequently.	89.7	82.8	75.9	89.7
089-Ability to recall information regarding specific events & activities (for example, tactics used in prior fire scenes).	86.2	79.3	72.4	82.8

*Note: All abilities were rated as important or critical by 2/3 of SMEs, except those marked in grey. *Exercises A and B were retained in the final version of the Firefighter CBT (Form A) based on subsequent validation studies.*

Summary

The new draft Firefighter CBT (Form A) was reviewed by a sample of FDNY Firefighters and Officers for the purpose of establishing the content validity of the Training Simulation (cognitive) portion of the Firefighter test. The session also afforded an opportunity for the Firefighters and Officers to pilot the CBT to ensure that the instructions and testing procedures were clear and practical, and to identify any needed refinements to the CBT prior to the criterion-related validation study. The content validation procedure yielded results supporting the job relevance of the new Firefighter test, in that the test exercises were linked to at least 14 of the 18 core Firefighter abilities that were identified in the 2011 FDNY Firefighter job analysis.

Criterion-Related Validity and Construct Validity Study

Evidence of criterion-related validity for the Firefighter CBT was established in an empirical study that demonstrated significant correlations between Firefighters' scores on the CBT and their performance in both the probationary academy and on the job. Evidence of construct validity was obtained in the same study by showing that FDNY Firefighters' CBT scores were significantly correlated with corresponding tests well established in the research literature as measuring the same or similar abilities and characteristics as those targeted by the CBT..

The study entailed a series of steps, including: (a) selection of a sample of FDNY Firefighters to take the CBT (Form A); (b) collection of job performance evaluations in research sessions conducted with the supervisors of the Firefighters who would be tested; (c) collection of available probationary academy performance records for Firefighters who would be tested; (d) administration of the CBT and other tests of the same abilities and characteristics to the sample of FDNY Firefighters; and (e) conducting an analysis of CBT items and test components to assemble a final form that is reliable, fair, and significantly correlated with Firefighter performance in the probationary academy and on the job.

Date(s), Time(s) and Location(s): The study was completed between May and November 2011. The data collection locations included the FDNY Academy and DCAS testing centers located in Manhattan and Brooklyn.

Sampling Plan

A sampling plan was established to select FDNY Firefighters who would complete CBT Form A, and for whom job performance data would be collected. The sample focused exclusively on full-duty Firefighters who were assigned to ladder or engine companies in the five boroughs of New York. The sample was designed to provide adequate statistical power to detect moderate effect sizes for the total sample and to facilitate subgroup analyses. To this end, a target sample of 811 incumbent Firefighters, including 250 each of Black, Hispanic and White incumbents, was randomly selected by PSI from an electronic file containing a list of all FDNY Firefighters.[23] The participants were randomly selected within the sampling strata, with over-representation of some groups to facilitate statistical analyses. Table 10 summarizes the validation study sampling plan, along with the total population of FDNY Firefighters.

[23] A statistical power analysis was conducted in establishing the target sample size, as follows: Assume the following to estimate power: r (null) = 0; r (true in applicant population) = .20; r (true in incumbent population, after 40% range restriction) = .12; N = 811; P < .05 (one tailed).

Table 10. Validation Study Sampling Plan

	Asian		Black		Hispanic		Native American		White		Male		Female		Total
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N
Validation Target Sample	64	7.90%	244	30.10%	250	30.80%	3	0.40%	250	30.80%	791	97.53%	20	2.47%	811
FDNY Firefighter Population	76	0.90%	312	3.70%	639	7.60%	5	0.10%	7342	87.70%	8350	99.70%	24	0.30%	8374

Job Performance Data Collection, File Construction and Quality Control

Job Performance Rating Sessions. PSI worked with the FDNY to identify supervisors of the 811 Firefighters selected for the study, who would be asked to complete the research-only performance evaluation. PSI requested the name of the officer who had completed the most recent evaluation of each validation study Firefighter so that individual could be scheduled to complete the research-only Job Performance Rating Booklet. In addition, for a subset of Firefighters, PSI requested that a second evaluator be identified complete a performance evaluation to enable the computation of inter-rater reliability estimates for the performance ratings. The Fire Department identified over 400 Fire Lieutenants and Captains and scheduled them to attend one of 11 performance rating sessions.

Prior to the rating sessions, the Special Master hosted a meeting with the parties' experts and leaders of the Firefighter and Officer unions to obtain their support for the project with assurances that the data would remain confidential and be used for research purposes only. The unions issued a communication to members encouraging their full cooperation in the project.

PSI conducted the rating sessions at the FDNY Training Academy on June 14-16, 2011. Make-up sessions were conducted on July 20, 2011. The DOJ's experts assisted in conducting some of the sessions, and the Vulcan Society and Special Master experts observed a number of the sessions.

The sessions were conducted as follows: A plenary session was conducted wherein all of the participants for the session assembled in the auditorium. During the plenary session, an overview of the project and purpose of the session was provided by PSI. Then participants were divided into subgroups of 10 to 15 officers to complete the ratings. During these breakout sessions, PSI and DOJ experts led the officers through the rating process, reviewing and discussing each part of the rating booklet, instructions and examples. Before completing each rating scale, the session facilitators and officers discussed the definition of the scale and gave examples of "good" and "poor" performance for the dimension. At the conclusion of the rating sessions, the officers provided their rating forms to PSI for review, prior to checking out. If incomplete forms were observed, the officers were directed to complete them prior to leaving.

Probationary Firefighter Training Academy Data Collection. PSI also obtained Firefighter Training Academy records for the sampled Firefighters from the FDNY academy

during June and July 2011. Data, including academy test scores, practical exercise scores, or both, were available for several hundred Firefighters in academy classes held between 2000 and 2008. These records were provided either in an electronic data file (Excel format), or via printed paper records of academy performance (which PSI staff key-entered into an electronic file).

Data File Construction and Quality Control. An electronic database was constructed and a preliminary analysis of the job performance data was conducted to ensure the quality of the data for purposes of the criterion-related validation study. PSI key-entered and verified the job performance ratings from the booklets and created an electronic data file. The data were subjected to a series of quality control checks (e.g., missing data, ratings of only "slightly familiar" with the Firefighter's performance, consistently extreme ratings of Firefighters). Appendix R outlines these quality control checks.

As a result of these analyses, useable ratings were retained for up to 755 Firefighters, depending upon the particular performance area being rated.

The electronic files containing the academy training records were merged with the job performance ratings to create a master criterion data file. As a result of these analyses, useable training data (academy test scores, practical exercise scores, or both) were obtained for 598 Firefighters.

Table 11 summarizes the sample of Firefighters for whom job performance ratings and academy data were obtained, with respect to borough, assignment, race/ethnicity, and gender. Because missing data vary across the various job performance and training measures, sample characteristics are provided for Firefighters for whom overall job performance ratings were obtained, and for whom midterm academy test scores were obtained.

Table 11. Characteristics of Firefighter Job Performance Sample

	Job Performance Ratings		Probationary Academy Data	
Firefighter Characteristic	**N**	**Percent**	**N**	**Percent**
Borough				
Brooklyn	193	26.0	153	26.2
Bronx	170	22.9	137	23.4
Manhattan	142	19.2	110	18.8
Queens	166	22.4	131	22.4
Staten Island	67	9.0	51	8.7
Other	3	0.0	2	0.3
Assignment				
Engine	463	62.5	375	64.1
Ladder	278	37.5	210	35.9
Race/Ethnicity				
Asian	58	7.8	46	7.9
Black	216	29.1	184	31.5
Hispanic	227	30.6	187	32.0
Native American	3	0.4	2	0.3
White	237	32.0	166	28.4

Gender				
Female	19	2.6	15	2.6
Male	722	97.4	570	97.4
Total	741	100	585	100

Note: For job performance, N is based on overall job performance rating; for academy data, N is based on midterm test score.

Job Performance Data Analyses

Performance Ratings. Descriptive statistics for the job performance ratings were computed including the mean, standard deviation (SD), and the number of ratings (N). The mean ratings of performance in the 18 Work Areas and Overall Effectiveness fell between "Fully Acceptable" (4) and "Very Good" (5). The mean rating of Firefighters on the five relative ranking scales fell between "Top 50% and "Top 30%", ratings of 4 and 5 respectively. The SD vales on the rating scales were generally between 0.8 and 1 scale point value, with few exceptions, indicating that there was substantial variation in Firefighter performance on the rating scales. Appendix S contains these descriptive statistics.

Estimates of rater reliability were computed by correlating the multiple ratings for a subsample of Firefighters who were each rated by two or more Officers (N= 79 to 169). In general, the rater correlations tended to be low (less than .30), which is likely due to the fact that the Firefighter's immediate supervisor was more familiar with the Firefighter's job performance than the second rater, who was asked to complete an additional rating for purposes of this study. In view of this finding, the single best rater per Firefighter was selected on the basis of his/her rating of familiarity with the Firefighter's performance (i.e., the most familiar rater was used rather than averaging the ratings of multiple raters for an individual Firefighter). Ratings on three scales were excluded from the study in view of their very low inter-rater reliability estimates (5. Taking up salvage; 7. Inspection; and 17. Communication).

Rating Composites. Composites of the job performance rating scales were constructed by grouping the scales on the basis of logical and statistical procedures (Principal Components analysis). The ratings were averaged for each group of rating scales. The results supported two task rating composites: (1) Fire Scene Activities, and (2) Station, Training and Medical Activities; as well as three ability/characteristic composites: (1) Problem Solving, Learning and Concentration, (2) Cooperation and Compliance and (3) Physical Performance.

Job performance rating composites were computed by averaging the ratings assigned to each of the five groups of rating scales. Descriptive statistics (means and SDs) for the resulting rating composites are reported in Table 12. The Principal Component factor loadings are shown in Appendix T.

Table 12. Descriptive Statistics for Job Performance Rating Composites

Rating Composite	N	Mean	SD
Fire Scene Activities	749	4.27	0.81
Station, Training & Medical Activities	756	4.42	0.77
Cooperation & Compliance	757	4.45	0.80
Problem Solving, Learning & Concentration	756	4.29	0.80
Physical Performance	756	4.66	0.91

Probationary Training Academy Performance. Descriptive statistics for Firefighters' probationary academy scores were computed, including the mean, standard deviation (SD), the number of Firefighters in the sample (N), and correlations among the academy scores. The quizzes, midterm and final exams were observed to be positively correlated (ranging from .17 to .36); correlations with practical exercise scores were mixed (ranging from -.15 to .37). Appendix U contains these descriptive statistics.

Correlations between academy performance scores and job performance are reported in Table 13. These results indicate that performance in the academy is statistically significantly correlated with job performance, particularly Problem Solving, Learning & Concentration; Station, Training & Medical Activities; Overall Job Performance; and Overall Performance Ranking.

Table 13. Correlations between Probationary Training Academy Scores and Job Performance Ratings

	Academy Performance			
Job Performance	Quizzes	Midterm	Final	Exercises
Effectiveness Ratings:				
Fire Scene Activities	.21**	.10*	.17**	.15*
Station, Training & Medical Activities	.18**	.15**	.09*	.01
Problem Solving, Learning, Con.	.22**	.10*	.15**	.13
Cooperation & Compliance	.10*	.11*	.04	.05
Physical Performance	.05	.05	.00	.02
Performance Complaints	.15**	.16**	.09*	.15*
Overall Performance Rating	.18**	.10*	.12**	.13
Relative Ranking:				
Technical Firefighter Capabilities	.19**	.06	.16**	.09
Getting Along with Others	.10*	.07	.12**	.05
Dependable Organizational Citizen	.18**	.12**	.11*	.06
Stamina & Strength	-.02	.00	.01	-.04
Overall Ranking	.17**	.12**	.11*	.06
No. Firefighters	539 – 550	538 - 549	501 – 512	208 - 212

***p<.01; *p<.05, two-tailed.*

Test Administration, Data File Construction and Quality Control

Scheduling and Preparation. During September 14-17, 2011 a total of 16 testing sessions were conducted at DCAS testing centers in Manhattan and Brooklyn, in which incumbent Firefighters selected for the validation study completed the experimental version of the CBT (Form A). PSI worked with the FDNY to schedule the incumbent Firefighters to attend the sessions. The Firefighters received advance letters from the Fire Commissioner and the Firefighter's union describing the project, its importance, and the need for confidentiality of the test materials and content.

Prior to the testing sessions, PSI staff worked with DCAS IT staff to configure and test the computers in the Manhattan and Brooklyn testing centers to ensure that the CBT software was functioning properly.

Testing Sessions. The testing sessions were supervised by PSI and were proctored by PSI and DCAS staff; IT staff also attended the sessions to provide technical assistance, if needed. FDNY Officers were on site to check-in the Firefighters upon their arrival at the testing center. PSI proctors then assigned the Firefighters to a computer workstation/seat and provided an instruction sheet and a test notebook (Training Guide), and escorted the Firefighters into the testing room. The Firefighters then started the CBT and proceeded to follow the instructions, which included a CBT tutorial and practice questions, and then completed experimental Form A. During the test, proctors actively walked around the testing room to ensure no one was talking or attempting to record the test content. The Firefighters completed the CBT at an individual pace and when finished, checked-out with the proctor who collected all notebooks and instructions. There were no incidents of unusual or irregular behavior reported by the test proctors, and no incidents of CBT technology failure or anomalies reported. A total of 725 Firefighters participated in the testing sessions.

Database Construction and Quality Control. The Firefighter test data were extracted from the CBT system to create an electronic file for purposes of analysis. The test data were subjected to a number of quality control steps that were applied to each portion of the test and cases were excluded from the analysis if they failed to meet certain criteria for reasonableness. Appendix V outlines the test data quality control criteria and data exclusion rules (e.g., omitted too many items; did not spend a reasonable amount of time; scored below chance-random guessing level).

As a result of these quality control checks, sample sizes for the various portions of the test ranged from 682 to 718; and 612 Firefighters had complete data across all test portions.

Test Data Analyses

Firefighter Examinee Sample Characteristics. A diverse sample of Firefighters completed the CBT. Table 14 summarizes the Firefighter sample, with respect to borough, assignment, race/ethnicity, and gender. Because the number of cases varied across test portions, the sample of 612 examinees with complete test data is reported in the table.

Table 14. Characteristics of Firefighters Tested in Criterion-related Validation Study

Firefighter Characteristic	N	Percent
Borough		
Brooklyn	142	23.2
Bronx	166	27.1
Manhattan	107	17.5
Queens	140	22.9
Staten Island	55	9.0
(not reported)	2	0.3
Assignment		
Engine	378	61.8
Ladder	234	38.2
Race/Ethnicity		
Asian	45	7.4
Black	187	30.6
Hispanic	185	30.2
Native American	3	0.5
White	192	31.4
Gender		
Female	17	2.8
Male	595	97.2
Total	612	100

Statistical Item and Test Analyses. A series of statistical analyses were conducted examining psychometric properties of the test items and test portions. The purpose of the analyses was to identify and eliminate any poorly functioning items, and to examine the properties of the various test components.

Classical item analysis procedures (Millman & Greene, 1989; Allen & Yen, 1979) were utilized to analyze examinees' responses to the experimental test items for purposes of calibrating psychometric characteristics (e.g., difficulty, item-total correlation, and distractor effectiveness). These item statistics provided a basis for selecting the best functioning items and eliminating poorly functioning items. The cognitive items and the noncognitive test items were analyzed differently, though the underlying statistics were similar.

Cognitive Test Items. An item analysis of the cognitive test items was conducted focusing on three psychometric properties:

1. Difficulty - the percentage of examinees responding correctly to an item (p);
2. Item-total Correlation - the item-total score point-biserial correlation (r), adjusted for spuriousness due to inclusion of the item in the total score (Henrysson, 1971); and
3. Distractor effectiveness - the item response-total score point-biserial correlation (r_d) and percentage of examinees choosing the alternative (p_d).

Cognitive test items were excluded from further consideration for inclusion in a test form if any of the following criteria were met:

- Low item-total correlations: adjusted r < .10;
- Extremely easy or difficult: p >95 or < 1/a * 100, where a=number of response alternatives;
- Ineffective distractor: positive r_d ($\geq$.05) in conjunction with substantial p_d ($\geq$10).

As a result of this analysis, a total of 9 cognitive items were eliminated and 143 were retained for further analysis in developing the new test.

Noncognitive Items. A similar item analysis was conducted for the noncognitive test items, in this case focusing on three psychometric characteristics:

1. Item Mean - the mean of scored item responses;
2. Response Variance - percentage of examinees choosing a single response option (p_o); and
3. Item-scale correlation - the item-scale score point-biserial correlation (r), adjusted for spuriousness due to inclusion of the item in the scale score.

In addition, an analysis of item linearity was conducted, reflected by a pattern of response option-scale correlations (r_o), where r_o is expected to increase monotonically from negative to positive across the ordered set of 5 response options. To achieve linearity of the ordered response options with biodata items, PSI collapsed response codes as necessary in an iterative process, until linearity was achieved

Noncognitive items were selected using the following process:

- Items with low discrimination values (adjusted r < .10) for the targeted scale were excluded
- Items with slightly higher discrimination values (adjusted r < .20) for the targeted scale were also considered for exclusion if other indicators (e.g., item mean, response distributions, alpha if item deleted) suggested potential problems with the item or with scale placement
- Using an iterative process, PSI then correlated deleted items with non-targeted scales to determine if items would be better placed in alternative scales.
- PSI conferred with DOJ's consultants on the final placement of items, and reclassified items based on rational and literature-based considerations, as well as item statistics.
- Likert-type items (i.e., where the response options included Strongly Disagree, Disagree, Agree, and Strongly Agree) were dichotomized to help minimize any effects of coaching on test scores in the subsequent applicant sample.

A total of 18 noncognitive items were excluded due to the above criteria, leaving 136 items from which to cast the 14 scales that were utilized in subsequent validation analyses.

Test Components. Descriptive statistics summarizing properties of the cognitive test components and the noncognitive component scales are summarized in Appendix W, including the mean, SD minimum and maximum score, ceiling index (the number of SDs separating the mean and maximum possible score), and reliability (alpha).

Criterion-related Validity Evidence

Sample Characteristics. The criterion-related validation sample is described in Table 15, reflecting Firefighters in the data file with test scores, job performance data and training academy performance data.

Table 15. Characteristics of Firefighters in the Criterion-related Validation Sample

	Test & Job Performance		Test & Academy Performance	
Firefighter Characteristic	**N**	**Percent**	**N**	**Percent**
Borough				
Brooklyn	130	23.0	106	23.8
Bronx	147	26.0	109	24.4
Manhattan	101	17.8	82	18.4
Queens	132	23.3	105	23.5
Staten Island	55	9.7	42	9.4
Other	1	0.2	2	0.4
Assignment				
Engine	352	62.2	282	63.2
Ladder	214	37.8	164	36.8
Race/Ethnicity				
Asian	42	7.4	33	7.4
Black	166	29.3	138	30.9
Hispanic	173	30.6	146	32.7
Native American	3	0.5	2	0.4
White	182	32.2	127	28.5
Gender				
Female	16	2.8	12	2.7
Male	550	97.2	434	97.3
Total	566	100	446	100

Note: N for Test & Job Performance is based on sample with usable test data across all test portions and overall performance rating present; N for Test & Training Performance is based on sample with usable test data across all test portions and midterm test score present.

Validity of Experimental CBT (Form A). The cognitive portion of the experimental CBT was comprised of 10 test components; the noncognitive portion was comprised of 14 test components (or scales). Product-moment correlation (validity) coefficients were computed between Firefighters' scores on various parts of the experimental CBT (Form A) and the measures of job performance and academy performance to examine the criterion-related validity of the test.

The results indicated that all of the cognitive test components were significantly correlated with performance in the Firefighter academy (significant correlations ranged from .09 to .40, corrected for criterion unreliability);[24] and nine test components predicted at least one job performance composite (significant correlations ranged from .11 to .26, corrected for criterion unreliability).[25] Seven of the noncognitive components predicted at least one job performance rating (significant correlations ranged from =.11 to .26, corrected for criterion unreliability).

Appendix X, Tables X-1 and X-2, report the validity coefficients obtained for the various cognitive components and noncognitive scales comprising the experimental CBT (Form A).

Assembly of Final CBT. The final CBT (Form A) was assembled in consideration of several factors, including: (a) maximizing validity; (b) representing important test content areas that are linked to core Firefighter abilities and characteristics; (c) minimizing racial/ethnic and gender group score differences; (d) yielding reliable scores with sufficient variance to be useful as a selection tool, and (e) practical administration time. To this end, the cognitive and noncognitive test components and scales were examined and selected for the final version for Form A, as described below.

The validity results for the cognitive test components were examined to identify a subset that would address the above considerations. The testing experts selected a 3-part video learning exercise and operations manual. Similarly, the experts reviewed the noncognitive scale validity results in conjunction with the core Firefighter characteristics and noncognitive assessment literature, and identified six scales for inclusion in the CBT: Agreeableness, Dependability, Even Tempered, Low Anxiety, Self Esteem, and Activity.

Criterion-related validity evidence is reported in Table 16 for the cognitive portion of the CBT, the noncognitive portion, and the CBT Total score (applying final scoring weights that are explained in Chapter 6). For summary purposes the cognitive job performance rating scales were averaged to create a cognitive performance composite; the noncognitive rating scales were similarly averaged. The academy criterion composite is the average of quizzes, midterm and final exam. Part "a" of the table reports validity coefficients corrected for unreliability, which for the final CBT total score ranged from .24 (predicting Overall Job Performance Ranking) to .30 (predicting Academy Performance). To better estimate the true validity of the CBT, an additional adjustment was applied for restriction in the range of test scores in the criterion-related study relative to the range of candidate scores, as shown in Part "b" of the table. The resulting validity coefficients for the final CBT total score ranged from .32 (predicting Overall Job Performance Ranking) to .39 (predicting Academy Performance).[26]

---

[24] The reliability of the Academy Performance composite was assumed to be .80, per the meta-analytic value suggested by Hunter and Hunter (1984).
[25] The reliability of the performance ratings ranged from .21 to .37 based on the criterion-related study.
[26] The CBT total score SD for candidates was 15.94; the SD for Firefighters in the validation study was 11.74.

Table 16. Criterion-related Validity Evidence for the CBT (Form A)

a. Validity Corrected for Criterion Unreliability

	Overall Job Performance Rating		*Overall Job Performance Ranking*		*Cognitive Performance Composite*		*Noncognitive Performance Composite*		*Academy Criterion Composite*	
	N	r	N	r	N	r	N	r	N	r
Cognitive score	601	.23**	604	.24**	615	.27**	615	.17**	472	.40**
Noncognitive score	650	.13*	653	.06	664	.12*	664	.19**	513	.03
CBT Total Score	592	.25**	595	.24**	606	.27**	606	.26**	464	.30**

***p<.01; *p<.05, one-tailed. Statistically significant correlations are corrected for criterion unreliability.*

b. Validity Corrected for Criterion Unreliability and Restriction in the Range of CBT Scores

	Overall Job Performance Rating		*Overall Job Performance Ranking*		*Cognitive Performance Composite*		*Noncognitive Performance Composite*		*Academy Criterion Composite*	
	N	r	N	r	N	r	N	r	N	r
Cognitive score	601	.29**	604	.30**	615	.34**	615	.21**	472	.49**
Noncognitive score	650	.15*	653	.06	664	.14*	664	.22**	513	.03
CBT Total Score	592	.33**	595	.32**	606	.36**	606	.34**	464	.39**

***p<.01; *p<.05, one-tailed. Statistically significant correlations are corrected for criterion unreliability and CBT score range restriction.*

Construct Validity Evidence

Evidence of construct validity was examined by correlating the CBT cognitive components and noncognitive scales with other previously published (marker) tests that have been demonstrated to measure the same abilities and characteristics (also referred to as constructs) intended to be measured by the CBT. The *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) recognize validity evidence based on the relationship between scores on tests hypothesized to measure the same thing.

> "Evidence that two measures are highly related and consistent with the underlying construct can provide convergent evidence of validity for the selection procedure. Evidence that test scores relate differently to distinct constructs can provide evidence of discriminant validity. "

Correlations were examined between the experimental CBT cognitive components and four well-established measures of cognitive ability from the Employee Aptitude Survey (EAS) test series (Ruch, et al. 2001). The EAS is itself supported by substantial evidence of criterion-related validity and construct validity. Overall, the pattern of correlations reflects the integrated nature of the CBT format, as the correlations with marker tests did not isolate specific constructs. The cognitive portion of the CBT entails a combination of observing, listening, reading, and applying detailed information which in some cases involves basic arithmetic. Appendix Y, Table Y-1

provides the matrix of test correlations for the final CBT cognitive components (ranging from .09 to .30).

Correlations were examined between the experimental CBT noncognitive scales and three well-established measures of workplace personality characteristics, including the Big-Five Inventory (BFI; John, 1991), the California Personality Inventory (CPI; Gough, 1996); and the Personnel Reaction Blank (PRB; Gough, 1971). Appendix Y, Tables Y-2 and Y-3 report the resulting correlations. The results indicated that, in general, the six scales selected for the final CBT were significantly correlated with corresponding personality test scores in a manner that would be expected. The results were as follows:

- CBT Activity scale correlated highest with the BFI Conscientiousness scale (.35);
- CBT Agreeableness scales correlated highest with BFI Agreeableness (.47);
- CBT Dependability scale correlated highest with BFI Conscientiousness (.55);
- CBT Even Tempered scale correlated highest with BFI Neuroticism (-.39);[27]
- CBT Low Anxiety correlated highest with BFI Neuroticism (-.54); and
- CBT Self Esteem correlated highest with BFI Conscientiousness (.39).

Investigation of Fairness[28]

An analysis was conducted to examine the predictive relationship between test scores and job performance for racial/ethnic groups (Blacks and Hispanics) to ensure that the tests were fair; i.e., did not substantially under-predict actual job performance. The analysis was consistent with the definition of fairness espoused in the *Uniform Guidelines* [Sec. 14.B.(8)(a)], which define unfairness as follows:

> "When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores."

To this end, a linear regression equation was derived for CBT scores relating to job performance rating composites for the Firefighters in the criterion-related validation study. Then "predicted" job performance ratings were compared to the actual ratings to compute a residual job performance score for Blacks and Hispanics relative to Whites (residual=actual minus predicted job performance). A mean residual score near zero would indicate that test scores are well representative of job performance for that group; a substantial *positive* mean difference would indicate *under*-prediction (potential unfairness); and a significant *negative* mean difference would reflect *over*-prediction; i.e., that the test is not unfair to that group.

---

[27] Negative correlation coefficients indicate an inverse relationship between the noncognitive scale score and the marker test score, as expected. For example, one would expect a highly "even tempered person" to attain a low score on "neuroticism," producing a negative correlation between the two measures.

[28] Uniform Guidelines Section 14 B.(8).

Results of the fairness analysis are summarized in Table 17. The mean difference scores (residuals) for the cognitive, noncognitive and Total CBT scores indicated that the CBT does not unfairly underestimate job performance for Blacks and Hispanics. This finding is consistent with generally reported findings in the scientific testing literature (SIOP, 2003, p. 32).

Table 17. CBT Fairness Analysis Results

		Black – White			Hispanic - White		
	Criterion	Residual Black	Intercept (t)	Slope (t)	Residual Hispanic	Intercept (t)	Slope (t)
Noncognitive Score	Overall Performance Rating	-0.18*	-3.91**	-0.54 NS	-0.15*	-3.36**	-1.44 NS
	Non-cog performance rating composite	-0.13	-2.98**	-0.80 NS	-0.06 NS	-1.37 NS	-1.83 NS
Cognitive Score	Overall Performance Rating	-0.17*	-3.59**	-0.63 NS	-0.13*	-2.98**	0.58 NS
	Cognitive performance rating composite	-0.21**	-5.27**	-1.26 NS	-0.14*	-3.70**	-0.60 NS
Total CBT	Overall Performance Rating	-0.16*	-3.43**	-0.61 NS	-0.14*	-3.07**	-0.61 NS
	Cognitive+ Non-cog performance rating composite	-0.32*	-4.09**	-0.63 NS	-0.19 NS	-2.59*	-1.37 NS

***p<.01; *p<.05, two-tailed. N = 363 – 403; results reflect final CBT algorithms and weights used for candidate scoring*

Summary

Evidence supporting the validity of Firefighter CBT scores was presented in three studies representing content, criterion-related and construct validation strategies. Evidence of content validity was established for the cognitive portion of the CBT by collecting data showing that the content of the test represents important aspects of performance on the FDNY Firefighter job.

Criterion-related validity evidence was established by demonstrating that incumbent Firefighters' CBT scores were significantly correlated with their performance in the Firefighter academy, as well as with Officers' ratings of their performance on the job. These results were used to select components of the experimental Form A that are correlated with academy and job performance, reliable, and practical to administer.

Construct validity evidence was established by demonstrating that noncognitive scores on the CBT are significantly correlated with other established tests of the same characteristics.

An investigation of test fairness indicated that the CBT was not unfair to Blacks and Hispanics in predicting their job performance.

CHAPTER 5: DEVELOPMENT OF ALTERNATE TEST FORMS

Introduction

In light of the large number of job applicants (approximately 60,000) who would sign up to take the Firefighter test, additional alternate CBT forms were developed to help safeguard the security and integrity of the examination; i.e., to minimize the exposure of the test items to candidates and create a barrier to cheating. This chapter describes the test development work and subsequent study that was conducted to assemble the alternate forms and document their equivalence to the validated Form A of the CBT.

Alternate Test Forms Development

Item Development

Alternate CBT forms were developed to be parallel to the original CBT-Form A, which was validated in several studies. The alternate forms contained unique cognitive portions and used the same noncognitive items. The rationale for using the same noncognitive items on the alternate forms was that they are difficult to clone and are not subject to the same types of cheating concerns as cognitive items.

To ensure that the alternate CBT forms would be equivalent to the original validated CBT Form A, items were developed following a cloning process wherein the Form A training lesson scripts were modified to create three alternate versions (B, C and D). These three variant scripts presented the same fictional device, with the same number of controls and gauges. In each alternate version, the device controls and gauges were renamed using fictitious labels of similar length to Form A, their functions were changed and their locations on the device were moved. These cloned scripts served as a basis for the development of alternate forms of the cognitive portions of Form A.

PSI consultants worked from three sets of scripts (Forms B, C, and D) to clone the test items contained in Form A, as well, using the new terminology and functions referenced in each alternative set of scripts above. In each case, the same item type was carried forward in the cloned item to match the original item type. These included multiple-choice single answer; multiple-choice multiple-answer; graphical item stems; graphical item response options; hot spot (click on a picture) items; and drag-and-drop items that required respondents to click-on and move things to indicate the appropriate order. In virtually every case, items were written as exactly the same type when cloning an item for the alternate forms. For the simulated operations manual (reading comprehension) portion, the reading passages were cloned from Form A and terms were substituted referencing different names and functions for controls and gauges. The reading level of the passages on the alternate forms was comparable across alternate forms (Flesch-Kincaid calibrations of the exam reading passages ranged from grade 9.2 to 11.9, and differed by 0.4 or less for parallel passages across forms).

An alternate version of the Training Guide was developed for each alternate form B, C and D), again by cloning the original version used in Form A and substituting the terms, functions, and pictures of the new devices.

Three alternate sets of video lectures were filmed on November 17, 2011, using the same actors to play the roles of instructor and student in each set of videos. Post-production of the media files took place over the following two weeks. The test items and media files were authored into the CBT system and subjected to quality assurance (QA) reviews by PSI staff.

Review of Test Items

During the first week of December 2011, PSI published draft forms of the video lectures and the test items on a secure website for review by testing experts representing the DOJ, the Vulcan Society and the Special Master. Comments and suggestions by the experts were reviewed and incorporated and various edits to videos were made (e.g., adding pauses to a video; and editing presentation slides, video images and questions).

A formal review of the alternate CBT forms was conducted with the testing experts representing DCAS/PSI, DOJ, the Vulcan Society and the Special Master at PSI offices in Burbank, California on December 15 and 16, 2011. The objective of the session was for the experts to "sit for" each form of the CBT and review the instructions, videos, test items and Training Guide and to identify and discuss possible revisions. The experts recommended a number of edits and enhancements, such as adding more pauses to the videos; modifying and clarifying the instructions; clarifying certain images and exhibits associated with test questions; and miscellaneous edits to certain test questions.

PSI executed the modifications recommended by the parties' experts and made numerous additional edits and refinements to the test videos, instructions and questions. The videos were edited and certain narrator parts for the test instructions were re-recorded. The updated items and videos were posted to the secure web site to provide the testing experts an opportunity to review the revised content.

PSI test development staff coordinated with IT staff to make final edits to the test questions and media files in the CBT system in preparation for administration of the tests in an equivalency study. As a result, three alternate forms of CBT were assembled for administration in a study. Each form contained extra items to provide flexibility in selecting items to construct forms that would be equivalent to Form A.

Equivalency Study

This section summarizes a study to design alternate forms of the CBT (Forms B, C and D) that was conducted to pretest, assemble equivalent forms and document their equivalence to the validated CBT Form A. The study was undertaken because Form A of the CBT, having been exposed to more than 700 Firefighters during the criterion-related validation study, could not be considered a secure version of the test for administration to actual job candidates.

Date(s), Time(s) and Location(s): The study was conducted during January and March 2012 in Los Angeles, CA.

Study Design and Sampling Plan

The equivalency study was designed to enable statistical item analyses and assembly of parallel or equivalent forms, consistent with methods outlined in professional testing standards (AERA, APA, NCME, 1999, *Standards for Educational and Psychological Testing*).[29]

The study design called for a sample of 675 people from outside the state of New York to be recruited to complete experimental versions of the alternate CBT forms (B, C, or D), along with Form A.

The alternate forms were administered in counter-balanced order wherein Form A was given first or second to approximately half of the participants. Accordingly, six test form configurations were used in the study (AB, BA, AC, CA, AD and DA). In each configuration, the noncognitive items were administered in between the two cognitive forms.

Once the data were collected and reviewed, statistical properties of the test items were computed and items were selected to construct final alternate forms that are equivalent to the validated CBT Form A, with respect to the following criteria for between-form similarity:

- Matched content;
- Equal mean (difficulty) ;
- Equal standard deviation (scale of measurement);
- Highly correlated with the validation Form A (measure the same abilities); and
- Equivalence to Form A within race/ethnicity and gender groups.

The sampling plan called for a total of 675 participants. The study sample targeted people in the Los Angeles area representing a range of hourly blue collar jobs (e.g., construction, light industrial, medical technology), with a high school or some college completed (not highly degreed). Equal numbers of participants were sought for the most prevalent racial/ethnic groups in New York City (Black, Hispanic and White) to enable subgroup analyses.

Test Administration

Recruitment of Participants. PSI engaged a staffing services firm to source participants for the study. The participants were paid for their time and were offered an incentive to score well on the test (entry into a lottery for a gift card). The staffing services firm used focused recruiting in an attempt to recruit subjects in proportion to the target demographic characteristics; however, it was not possible to exactly match demographic targets as there were concerns by the staffing firm about violating employment laws if referrals were limited on the basis of age, race or gender.

[29] The *Standards* describe **alternate forms** as "Two or more versions of a test that are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. *Alternate forms* is a generic term used to refer to any of three categories. *Parallel forms* have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. *Equivalent forms* do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversion to derived scores or in form-specific norm tables. *Comparable forms* are highly similar in content, but the degree of statistical similarity has not been demonstrated."

Scheduling and Preparation. During January 2–29, 2012 a total of 46 testing sessions were conducted at four PSI testing centers located in the Los Angeles area, including Anaheim (2 sessions), Burbank (32 sessions), Carson (6 sessions), and El Monte (6 sessions). PSI worked jointly with the staffing firm to schedule participants for the sessions. The participants were informed that the test was a vocational assessment for blue collar jobs. There was no mention of FDNY Firefighters and all test materials referred to only as the "Vocational Test Study."

Testing Sessions. The test sessions were conducted by PSI test proctors who are experienced in the administration and supervision of high-stakes testing, and who ensured that the Firefighter CBT was administered in a secure and standardized manner. The proctors were responsible for participant check-in (verifying the participants with a picture ID vs. the roster provided by the staffing firm) and randomly assigning participants to a computer station and one of the six versions of the alternate CBT (AB, BA, AC, CA, AD, and DA). PSI proctors then provided an instruction sheet and a test notebook (Training Guide), and escorted the participants into the testing room. The participants then started the CBT and proceeded to follow the instructions, which included a CBT tutorial, followed by completion of their assigned test version. During the test administration, proctors actively walked around the testing room to ensure no one was talking or attempting to record the test content. The participants completed the CBT at their own pace and when finished, checked-out with the proctor who collected all notebooks and instructional materials. A total of 718 people participated in the equivalency testing sessions.

Database Development and Quality Control

The equivalency study data were extracted from the CBT system to create an electronic file. The test data were subjected to a quality control review that was applied to each form of the test and cases were excluded from the analysis if they failed to meet the criteria for reasonableness (see Appendix Z). Participant test data records were excluded from the analysis if they scored at or below chance-level on Form A and the alternate form taken (Form B, C or D); i.e., obtained a total cognitive score that would result from random guessing.

After application of these rules, the test items were checked to ensure that each item was answered by at least 90% of examinees. The results indicated that all items met this criterion, and thus no items were dropped from the analysis.

The resulting sample of participants who completed Form A and one of the Alternate forms (B, C or D) is summarized in Table 18 with respect to race/ethnicity, gender, education, age, order of tests given, and responses to career interest and experience questions. Because the number of cases varies across test form pairs, the sample of 688 examinees with complete test data is described in the table.

Overall, the equivalency sample represented the target sample characteristics per the sampling plan, with respect to race and gender. A departure from the target was that a portion of the examinees were over age 36. However, analyses of age group scores on the alternate CBT forms indicated that there was no significant difference between age groups on the Alternate CBT forms; thus, the data for people over age 36 were retained for the analysis.

Table 18. Characteristics of the Equivalency Study Sample

Participant Characteristic	N	Percent
Race/Ethnicity		
Asian	23	3.3
African American/Black	193	28.1
Hispanic	237	34.4
Native American	4	0.6
Caucasian/White	185	26.9
Other	44	6.4
Unknown	2	0.3
Gender		
Female	273	39.7
Male	415	60.3
Test Order		
AB	115	16.7
BA	115	16.7
AC	115	16.7
CA	109	15.8
AD	118	17.2
DA	116	16.9
Total	688	

Note: N for sample after application of data quality criteria.

Item Analyses

Similar to the procedure for assembling the original validated CBT Form A, a series of statistical analyses were conducted examining psychometric properties of the cognitive test items and portions comprising the alternate forms. The purpose of the analysis was to identify and eliminate any poorly functioning items, and to examine the properties of the various test portions for purposes of assembling final versions of alternate forms of the validated CBT Form A.

The item analysis of the alternate form cognitive test items was conducted focusing on the same psychometric properties as Form A (i.e., difficulty, item-total correlation, distractor effectiveness) and items were subjected to the same criteria for inclusion. (The noncognitive portion of the alternate forms used the same items as Form A). Table 19 summarizes the results of the item analysis for retained items by test form.

Table 19. CBT Cognitive Item Analysis Summary

	Form A	Form B1	Form B2	Form C	Form D
No. Items	56	57	57	57	57
Item Difficulty					
Mean	.67	.69	.68	.69	.69
SD	.18	.18	.17	.17	.19
Minimum	.10	.19	.19	.25	.19
Maximum	.90	.97	.97	.94	.95
Point-biserial correlation					
Mean	.42	.37	.39	.45	.45
SD	.09	.10	.10	.08	.10
Minimum	.18	.15	.19	.24	.20
Maximum	.62	.60	.56	.63	.65
N (range)	*666-687*	*221-230*	*220-230*	*214-224*	*212-234*

Note: Item difficulties expressed as proportion correct of available points per item; item analysis is based on equivalency study sample.

Alternate Form Assembly and Equivalence Analysis

Alternate forms were assembled from each of the three sets of items, resulting in three primary alternate forms (B1, C1 and D1) which had unique video-based lessons and training guides. Three additional forms were assembled (Forms B2, C2 and D2) using the remaining items with the same video lessons. However, only Form B2 was used; the other two forms (C2 and D2) were held in reserve and not used to test candidates. Form B2 item content overlapped with Form B1 item content by 46%.

Overall, the alternate CBT Forms were found to be highly correlated with the validated Form A ($r >= .87$), indicating that the forms provide comparable measurement of candidate abilities and characteristics. The forms were observed to be equivalent with respect to item content, raw score means and reliability. These results for the four operational forms are reported in Table 20.

The forms varied slightly with regard to standard deviations in the equivalency sample, indicating that a small scale adjustment would likely be required to locate examinee scores on exactly the same scale of measurement. This re-standardization was later conducted using the much larger candidate sample.

The results of the equivalency study were reviewed by the DOJ and Vulcan Society experts and used as a basis for determining to proceed with the project.

Table 20. Equivalence of Alternate CBT Forms

Form	No. Items	Max possible score	No. Examinees	Mean	SD	Alpha	Correlation with Form A
Validation Form A	121	128.4	695	79.2	21.7	.90	n/a
Form B1	122	128.4	237	78.7	20.5	.87	.87
Form B2	122	128.4	237	78.8	20.7	.89	.87
Form C1	122	128.4	233	78.1	24.5	.91	.89
Form D1	122	128.4	239	80.1	21.0	.89	.87

Note: Results reflect final CBT algorithms and weights used for candidate scoring, and standardized to candidate mean and SD.

Summary

Additional alternate CBT forms were developed for administration to candidates to help safeguard the security and integrity of the examination; i.e., to minimize the exposure of the test items to candidates and create a barrier to cheating. An equivalency study was conducted which enable the pretesting of the new items, assembly of equivalent forms, and documentation of their equivalence to the validated form of the CBT (Form A).

CHAPTER 6: SCORING AND USE OF THE FIREFIGHTER TEST

Introduction[30]

This chapter summarizes the method for scoring and use of the Firefighter CBT. The scoring procedures were developed by PSI prior to the administration of the CBT to job candidates. The method was derived using data gathered in the FDNY Firefighter job analysis, criterion-related validation, and alternate form equivalency studies described in earlier chapters. After the test was administered, a Test Validation Board identified scoring key modifications for certain items and those were incorporated into the final scoring. In addition, the final scoring procedure incorporated various mandated credits (bonus points) associated with City residence, veteran status, and legacy status; all credits established by the City well before design of the current test, and required by the New York State Constitution and laws, and longstanding City policy.

Scoring Procedure and Rationale

The scoring procedure included: (a) the assignment of weights to different portions of the CBT (weights determine the relative contribution of a portion to total CBT score); (b) application of a minimum passing score; (c) converting CBT scores to a 100-point integer scale; and (d) the application of mandated credits for veterans and disabled veterans, New York City residents, and candidates whose parent or sibling was a Firefighter killed in the line of duty (Legacy credit).

Test Component Weights

As noted earlier, the CBT was comprised of three major components: (1) Video Lesson: a 3-part video-based training lesson about a fictitious device, (2) Operations Manual: reading a simulated operations manual pertaining to the fictitious device, and (3) Background Survey: background questions covering six noncognitive areas. The first two components measure cognitive abilities and the third component measures noncognitive characteristics.

The 2011 FDNY Firefighter job analysis results and the 2011 criterion-related validity results were used to identify a job-related method for combining these components into a total CBT score.

Cognitive Test Component Weights. First, consideration was given to determining the relative weight to apply to the two cognitive components (Video Lesson and Operations Manual) in computing a total cognitive score (i.e., weighting within the cognitive total score). The weighting of the cognitive components relied upon the 2011 Firefighter job analysis data. Because the Video Lesson and Operations Manual were highly correlated (r=.78), it was not feasible to explore regression-based weighting (i.e., examining the relative size of the regression weights in predicting job performance ratings).

Scoring weights were derived for the cognitive test components by examining the relative importance of the abilities measured by each component, as indicated in the 2011 Firefighter job analysis. The relative importance of each ability was determined by summing the job analysis ratings for those Task Categories linked to the ability during the job analysis. More specifically,

[30] Uniform Guidelines Sec. 14 B.(6), C. (8); 15 B.(10, C.(7), D.(9).

during the Ability-Task Category linkage survey, Firefighters and Officers rated each Task Category on a 100-point scale with respect to its relative importance compared to the other Task Categories. For each test component (video, operation manual) the mean point value assigned to each linked Task Category was summed and then averaged across the abilities measured. The means of the linked Task ratings for the Video Lesson (55.0) and Operations Manual (25.2) were rescaled to relative percentages of 71 and 29, and then rounded to the nearest decile (to avoid over-precision in the deriving the weights), resulting in scoring weights of .70 and .30 for the Video Learning and Operations Manual components of the cognitive test, respectively.

Appendix AA, Table AA-1 summarizes the derivation of the scoring weights for the video lesson and operations manual components of the cognitive portion of the CBT. Table AA-2 in the appendix lists the corresponding abilities (KSAs) identified in the 2011 job analysis that the cognitive test components were designed to measure.

Cognitive and Noncognitive Portion Weights. Next, consideration was given to combining scores on the cognitive and noncognitive portions of the CBT. Alternative approaches to weighting cognitive and noncognitive tests in a compensatory total battery score were considered, including methods based on: (a) job analysis results, (b) regression of tests onto job performance ratings, and (c) equal weighting. Appendix AA, Table AA-3 summarizes the job analysis, regression-based, and equal weighting results.

The job analysis-based approach was similar to that described above for within-cognitive portion weighting, wherein the sums of linked Task Category ratings were averaged across the abilities and characteristics measured by the cognitive and noncognitive test portions. This approach yielded weights for cognitive total and noncognitive total scores of 45% and 55% respectively (45/55).

A regression-based weighting method was examined using data from the criterion-related validation study, wherein a multiple regression analysis was conducted with cognitive (70/30 weighted portions) and noncognitive total scores predicting job performance ratings (composite of cognitive and noncognitive performance rating scales). The resulting regression coefficients were rescaled to relative weights by dividing each weight by the sum of the weights. This approach yielded scoring weights of 64% cognitive and 36% noncognitive.

Because the job analytic yielded higher weight for the noncognitive potion and the regression-based approach yielded higher weight yielded higher weight for the cognitive portion, a moderate solution (roughly equivalent to averaging the two approaches) was adopted which gave equal (50/50) weight to the cognitive and noncognitive portions of the CBT.

Minimum Passing Score

A minimum passing score (cut score) was derived for the CBT to determine the minimally acceptable level of performance on the test, below which candidates are not likely to be successful on the job, and above which candidates are recommended for further consideration in the hiring process. The criterion-related validity (CRV) evidence supports making further distinctions between candidates who score above the cut score (i.e., use of scores for ranking).

The recommended cut score was derived drawing from the 2011 CRV study data by determining the score on the Total CBT that corresponds to minimally acceptable job performance among the

Firefighters who completed the CBT (Form A), and for whom special research-only job performance ratings were obtained. A statistical analysis (linear regression) was conducted wherein the empirical relationship between Total CBT score and mean job performance rating composite (mean cognitive + mean noncognitive rating) was quantified in a linear prediction equation. Using this linear equation, the Total CBT score (computed as described above) was identified which corresponds to minimally acceptable performance (a rating of "Just Adequate" or "3") on the job performance rating scale, adjusting downward for the standard error of estimate (SEE) of the regression equation. Appendix AB shows the results of the regression analysis to derive the minimum passing score. These results reflect updates that were made after the Test Validations Board convened and recommended certain item scoring modifications (described later in this Chapter).

Test Score Scale

An important consideration in test scoring is the scale upon which the scores will be reported to test takers and used for decision making. The scale refers to the unit of measurement, and includes features such as the range of point values, degree of score precision (number of decimal places used versus rounding, and the resultant intervals between scores), and the score distribution (e.g., mean and SD).

Several factors were considered to develop a recommended score scale for the CBT that would enable the City to distinguish among candidates for purposes of Firefighter selection. These included:

1. The CRV study results support the use of Total CBT scores to identify candidates who are most likely to be the better performers in the Firefighter Academy and on the job.
2. The precision of measurement associated with Total CBT scores was high (alpha = .88) and supported integer scoring units.
3. The City has historically reported the passing score as 70.
4. Bonus points must be added to the total test scores of persons who meet certain eligibility requirements (Veteran, Disabled Veteran, NYC resident, Legacy).

In light of the above considerations, Total CBT scores were converted to a 100-point integer scale. The scaling procedure uses two linear formulas: one for scores at or above the cut score to convert to a scale of 70 to 100; the other for scores below the cut score to convert to a scale of 0 to 69. The final CBT score conversion formulas derived for each form of the CBT (B1, B2, C1 and D1) are shown in Appendix AC.

To explore the reasonableness of the CBT 100-point integer scale prior to its use with candidates, it was applied to the Equivalency study sample and the score distribution was examined to confirm that the 100-point scales would provide a useful means of gauging candidate performance on the test. The 100-point integer scale was found to yield a fairly even distribution of scores that did not result in a large percentage of candidates obtaining the same score. Also, the 100-point scale enables adding bonus points in a straight-forward manner that could be readily explained to candidates. A summary of candidates' converted scores is provided in Appendix AD.

Mandated Bonus Points

New York State Constitution and laws and City policy provide bonus points to eligible candidates who request them. Table 21 outlines the available bonus point credits. The points are required to be added to passing examinees' total test scores. According to state law, eligibility for bonus points and the amount awarded are different for promotional candidates who are presently employed permanently by the City as Emergency Medical Specialist (EMT or Paramedic), than for candidates who are not permanently employed by the department (open pool).

Table 21. Bonus Point Credits

Points	Type of Credit
	Veterans Preference – based on NYS Constitution:
	Honorable discharge during time of war (as defined in NYS Law)
5	Open pool candidate
2.5	Promotional candidate
	Disabled Veteran
10	Open pool candidate
5	Promotional candidate
5	NYC Residency Credit – based on FDNY request for Exam No. 0084 and announced in August 1994 amended Notice of Exam: continuous NYC residency from July 1, 2010 through June 30, 2011 and is specified on the Notice of Exam (open pool candidates only)
10	Parent Legacy Credit – initially based on a policy decision, but now defined more narrowly in NYS Civil Service Law: for a parent who died in the line of duty as a Firefighter or police officer in the service of NYC (open pool candidates only)
10	Sibling Legacy Credit – based on NYS Civil Service Law: has a sibling who died in the line of duty as a Firefighter or police officer in the service of NYC as a result of September 11, 2001 WTC attack or rescue efforts in response to that attack. (open pool candidates only)

Note: Candidates may receive any combination of the above four credits, including points for both Parent Legacy Credit and Sibling Legacy Credit, yielding a possible total of up to 35 bonus points.

In summary, the CBT scoring procedure entailed the following steps:

1. Compute CBT cognitive score (70/30 weighting of video learning and operation manual); convert to T-score (mean=50, SD=10) rounded to 1 decimal;
2. Compute Total CBT score (50/50 weighting of cognitive and noncognitive portions), convert to T score rounded to 1 decimal;
3. Apply minimum passing score ;
4. Convert T scores to 100-point integer scale (rounded with no decimals) with passing score =70; and
5. Add bonus points to produce the Adjusted Final Score.

Tutorial for the CBT

PSI developed a tutorial for the CBT system and practice questions to help ensure that candidates were familiar with the test format. Candidates were notified of the availability of the tutorial in the exam registration letter that was disseminated several weeks before the test administration. The tutorial was posted on the DCAS website and was assessable via any computer with an Internet connection (e.g., from home, the library, DCAS offices, and Firehouses throughout the City).

The tutorial was comprised of a narrated slide show which contained information about the use of CBT system, with pictures of the computer screens, instructions and navigation windows and messages. The tutorial also contained sample test questions illustrating each type of question that was included in the Firefighter CBT, and included answers to the questions. General information was also provided regarding the content of the test.

Administration of the CBT

The CBT was administered to candidates between March 15, 2012 and August 1, 2012, at 15 testing centers located throughout New York City and several outlying areas. A total of 42,231 candidates were tested, of which 873 candidates were EMTs/Paramedics employed by the City who were eligible to take the promotional exam, classified as Exam 2500; and 41,358 were open competitive pool candidates who were classified as Exam 2000. [31]

The testing sessions were proctored by trained PSI and DCAS staff that actively monitored candidates during the testing sessions. During check-in, candidates placed their cell phones and all belonging into sealed plastic bags, which were then placed under their seats. Candidates were not allowed to talk during the test. Each computer station was separated by a carrel to prevent candidates from looking at other computers. The test items were delivered in random order within portions of the test and items were not numbered, so it would be extremely difficult for candidates to copy answers from other candidates.

Four forms of the CBT were administered to candidates: Forms B1, B2, C1 and D1. Forms B, B2, and C1 were used during the first week. Form D1 was introduced on April 1, 2012.

During the administration of the exam, candidate performance on each form of the test was monitored on a daily basis, by testing location. There was no evidence that scores were increasing over time, indicating that the exam content was not compromised.

Post-Administration Analysis

Confirmation of Scoring Key

After the CBT was administered to candidates, an item analysis was conducted to confirm that the test items were properly scored and to detect any unexpected problems with the items that

[31] The primary testing period ended April 20, 2012. However, the City is required to continue to provide make-up tests to a very small number of candidates who were away on active military duty. The data described in this section includes makeup sessions held through August 1, 2012.

were not identified in the criterion-related validation study and the equivalency study (e.g., mis-key, extreme difficulty). No problems were detected with the item statistics and, thus, no changes were made at the time.[32]

The results of the item analysis for each CBT form indicated that the test item key was appropriately applied and that the test items exhibited acceptable psychometric properties in the candidate sample (i.e., acceptable difficulty, positive item score correlation with total score, and distractor effectiveness). Appendix AE summarizes the results of the item analysis for the cognitive sections of the four alternate forms, as well as the noncognitive section (which was the same on each form). These item analysis results reflect the final scoring of the items, subsequent to application of the TVB decisions described in the next section.

Test Validation Board Scoring Adjustments

A Test Validation Board (TVB) was convened as required by NY Civil Service Law and court order, during the month of June 2012 to review and adjudicate item protests that were submitted by candidates during formal test review sessions conducted after the test administration. The TVB was comprised of three members, one chosen by the City, one chose by the Firefighter's union, and one chosen by the Court in response to a motion filed by the City. The purpose of the TVB was to review and adjudicate item protests submitted by candidates who attended protest review sessions in May 2012, after the test administration had been completed. The TVB recommended adjustments to the scoring of certain items, as follows: three (3) cognitive items were identified (two items on Form B2 and one item on Form D1) for which one additional response option in addition to the originally keyed response would be counted as "full credit"; and six (6) noncognitive items were identified for which full credit would be awarded for any of the response options. The TVB also awarded partial credit for six (6) additional noncognitive items.

In addition to the TVB modifications, two cognitive test items on Form D were given full credit for all examinees because of a computer delivery issue that resulted in several hundred candidates receiving extraneous information during the delivery of those items. While it was not apparent that this caused a problem, the fair action to take was to effectively neutralize these two items by giving all examinees credit.[33]

These item scoring modifications were enacted prior to computing candidates' final CBT scores. Subsequent analyses indicated that the scoring modifications did not affect the equivalence of the alternate forms, although small adjustments to the scale were needed (see below). Statistical analyses further showed that the TVB's changes to the scoring of the 12 noncognitive items did not have a significant effect on criterion-related validity for that portion of the test. While it was not possible to perform statistical analyses to determine whether the changes to the three

---

[32] Because of the complex nature of the scoring, PSI undertook an effort to ensure that all steps required to score exams were applied accurately within the SAS statistical package utilized to score candidates. The process involved the following steps: (a) Compute scores within SAS using the process and algorithms as described in this report; (b) Select one passing and one failing candidate from each form of the exam (B1, B2, C1 and D1) for verification and extract item response and scoring formula results for these 8 candidates; (c) Starting with unscored item responses, manually score each candidate's test results utilizing the appropriate scoring key and transformation formulas; (d) Compare the SAS-generated scoring results with the hand scoring results to ensure full correspondence. All interim scoring results for parts of the test and final results were reviewed and matched to at least 5 decimal places.

[33] One item was answered correctly by 99% of candidates; the other was answered correctly by 59% of candidates.

cognitive items impacted criterion-related validity (because incumbent Firefighters did not take those items in the validation study), the changes to the cognitive items were so small in number that it is unlikely they significantly affected the criterion-related validity. Furthermore, the reliability of the alternate forms was not affected by the above item scoring modifications, as the reliability coefficients (alphas) remained essentially the same in magnitude before and after application of the scoring modifications. See Appendix AF.

Candidate Final CBT Scores

CBT total scores were computed for candidates following the scoring procedure outlined earlier in this chapter. This procedure was applied separately for each CBT Form (B1, B2, C1 and D1) in order to further ensure equivalence of the forms.

The resulting candidate pass rates are summarized in Table 22, which shows results for the total candidate pool, and separately for the open competitive pool of candidate s (Exam 2000) and internal/promotional candidates (Exam 2500). The overall passing rate was 97.8%, and comparable for open pool candidates (97.7 %) and promotional candidates (99.5 %).

Table 22. Candidate Passing Rates on the CBT

	No. Candidates	No. Pass	Passing Rate
Internal Candidates (Exam 2500)	873	869	99.5%
Open Competitive Pool (Exam 2000)	41,358	40,426	97.7%
Total Candidate Pool	42,231	41,295	97.8%

Table 23 shows the distribution of final CBT scores (including bonus points) for the 869 internal promotional candidates (Exam 2500) who obtained a passing score, with scores ranging from 72 to 100.5. Table 24 shows the distribution of final CBT scores (including bonus points) for the 40,426 candidates in the open competitive pool (Exam 2000) who obtained a passing score, with scores ranging from 70 to 118. Overall, these tables indicate that the distribution of final CBT scores exhibited substantial variability among candidates, without truncating or "piling up" at the top of the score range.

Table 23. Distribution of Final CBT Scores for Promotional Candidates (Exam 2500)

CBT Final Score	No. Candidates	Cum. No. Candidates	Cumulative % Candidates
100.5	2	2	.002
99	7	9	.010
98	16	25	.029
97.5	1	26	.030
97	30	56	.064
96.5	1	57	.065
96	55	112	.128
95.5	3	115	.132

95	53	168	.192
94	60	228	.261
93.5	2	230	.263
93	82	312	.357
92.5	1	313	.359
92	89	402	.460
91	66	468	.536
90.5	2	470	.538
90	60	530	.607
89.5	1	531	.608
89	59	590	.676
88.5	1	591	.677
88	62	653	.748
87	49	702	.804
86	37	739	.847
85	22	761	.872
84	31	792	.907
83	22	814	.932
82	13	827	.947
81	9	836	.958
80	10	846	.969
79	6	852	.976
78	4	856	.981
77	2	858	.983
76	5	863	.989
75	2	865	.991
74	1	866	.992
73	1	867	.993
72	2	869	.995

Table 24. Distribution of Final CBT Scores for Open Competitive Pool of Candidates (Exam 2000)

CBT Final Score	No. Candidates	Cum. No. Candidates	Cumulative % Candidates
118	1	1	.000
116	1	2	.000
113	4	6	.000
112	8	14	.000
111	10	24	.001
110	7	31	.001
109	12	43	.001
108	23	66	.002
107	38	104	.003

106	51	155	.004
105	95	250	.006
104	122	372	.009
103	337	709	.017
102	595	1304	.032
101	856	2160	.052
100	1105	3265	.079
99	1443	4708	.114
98	1679	6387	.154
97	2161	8548	.207
96	2340	10888	.263
95	2425	13313	.322
94	2623	15936	.385
93	2728	18664	.451
92	2553	21217	.513
91	2479	23696	.573
90	2304	26000	.629
89	2079	28079	.679
88	2009	30088	.728
87	1687	31775	.768
86	1550	33325	.806
85	1301	34626	.837
84	1089	35715	.864
83	939	36654	.886
82	761	37415	.905
81	657	38072	.921
80	533	38605	.933
79	436	39041	.944
78	375	39416	.953
77	325	39741	.961
76	233	39974	.967
75	175	40149	.971
74	68	40217	.972
73	63	40280	.974
72	68	40348	.976
71	53	40401	.977
70	25	40426	.977

Use of CBT Scores to Select Candidates

The Firefighter CBT was developed and validated for use in assessing Firefighter candidates as a competitive examination (i.e., for use determining the order in which candidates would continue in the hiring process in descending order of score). DCAS has specified a manner of use of CBT scores in accordance with its established regulations, as described below.

Eligible Lists

DCAS will create two separate eligible lists of passing candidates, one for the promotional candidates (Exam 2500), and one for the open competitive pool of candidates (Exam 2000). DCAS will exhaust the promotional candidate eligible list before selecting open pool candidates. Both lists are "top down" rankings, based on final CBT scores after applying all bonus points claimed by the candidates. These CBT scores plus bonus points are called the "Adjusted Final Scores."[34]

The FDNY appoints a class of approximately 300 probationary Firefighters at a time to the Academy. In order to provide a final group of 300 candidates who successfully complete all remaining steps in the hiring process, the FDNY requests that DCAS certify 1,200 names to the FDNY at a time. However, because all candidates with tied Adjusted Final Scores must be certified at the same time, there may be instances when more than 1,200 candidates are certified to the FDNY at once. Typically, the FDNY will appoint two academy classes per year, and thus would hire 600 probationary Firefighters per year, requiring certification of at least 2,400 candidates from the eligible lists annually.

All candidates certified by DCAS to the FDNY are scheduled for CPAT (Candidate Physical Ability Test) orientation sessions (as required by the test license) and are administered the CPAT after an opportunity to train for at least nine weeks. Selected candidates also receive employment packets from FDNY's Candidate Investigation Division (CID) which schedules an intake interview. The field background investigation is then conducted and candidates undergo psychological and medical screening.

If more than 300 of the certified candidates successfully complete post-exam processing, the City will appoint candidates to the Academy based on the candidates' Adjusted Final Scores. Under this approach, the City may have to break tie scores between the candidates who successfully complete processing. For purposes of breaking ties between candidates with the same Adjusted Final Score, the City assigns candidates a list number, for administrative purposes only, that is based on the last five and then the first four digits of their social security numbers ("SSN"), with the higher SSNs receiving the lower list numbers. When ties must be broken, the City will appoint candidates based on this list number. The City will only deviate from this tie-breaking process in situations where a candidate has not completed processing by the date that the City begins to notify candidates of their appointment to the Academy, which is two to three weeks before the Academy class begins.[35] Thus, the City will not make any selection decisions among candidates with tied scores until two to three weeks before the Academy class begins. Importantly, the City will not skip over a certified candidate because he or she has not yet been processed (*i.e.*, received a disposition code) if that candidate has a higher Adjusted Final Score than other certified candidates who have successfully completed processing. Instead, the City will wait until the candidate has been processed (*i.e.*, receives a disposition code) before it appoints candidates with lower Adjusted Final Scores.

---

[34] Only candidates who obtained at least a score of 70 on the CBT are placed on the eligible list and receive bonus points.

[35] If the City skips over candidates because they have not completed processing, the City may still appoint them to the Academy if they complete processing before the Academy begins and open slots in the Academy still remain.

Successful candidates in excess of the 300 needed for the class, as well as candidates on the certification whose score is not reached are told to standby and are certified for the next class. At the time of the next class, the background investigation or the medical/psychological screening may need to be updated, but the CPAT does not need to be repeated. For the next class, after determining the number of candidates still available for hire from the earlier certification, DCAS will certify more candidates, in rank order, from the eligible list(s) to ensure that the FDNY has approximately 1,200 candidates to consider for each class. Since all candidates at the same score are certified at the same time there may be certifications that include more than 1,200 candidates.

Projected Selection Rates over the Life of the Eligible Lists

This section estimates candidate selection rates on an annual basis given: (a) the candidate results, (b) the number of Firefighter trainee positions available on an annual basis (600), (c) DCAS' above described procedures for establishing and selecting candidates from eligible lists, and (d) the number of the number of candidates historically selected from the eligible lists to account for attrition in subsequent parts of the selection process.

Table 25 reports projected selection rates over years, assuming: 600 position openings per year to fill two academy classes; and that a select/hire ratio of about 2 to 1 is needed for promotional candidates and about 4 to 1 for open competitive candidates (Exam 2000) to obtain a sufficient number who will successfully complete subsequent steps in the selection process (e.g., physical ability test, background investigation, medical exam). These select/hire ratios are what the City has typically experienced in order to obtain a sufficient number of candidates.

Applying these projections, a total of approximately 9,417 internal and open pool candidates (22.3%) would be expected to be selected from the eligibility lists over 4 years to fill 2,400 positions. These estimates may not reflect the actual number of candidates selected, because it is not possible to predict who will pass the subsequent steps in the section process, such as successfully completing the CPAT, background investigation, and medical and psychological exams.

Table 25. Projected Annual Candidate Selection Rates

Source	*No. Positions open*	*No. Candidates*	*Minimum Score Obtained*	*No. Selected to Proceed*	*Cumulative Pass Rate*	*Selection Ratio*
Internal Pool						
Year 1	600	873	72	869	0.995	1.4
Open Pool						
Year 2	600	41358	101	2160	0.052	3.6
Year 3	600	39198	99	2548	0.114	4.2
Year 4	600	36650	97	3840	0.207	6.4
Total 4 years	2400	42231	--	9417	0.223	3.9

Note: Selection Ratio = No. selected / No. Positions open. Cumulative Pass Rate is within pool; values for Total 4 years are based on combined pools.

Adverse Impact Analyses

Estimated selection (processing) rates from the two rank-ordered candidate lists were compared between racial/ethnic and gender groups and analyses were conducted to identify any potential adverse impact for both internal and open competitive pool candidates. Adverse impact was analyzed using the "four-fifths rule" (Uniform Guidelines Sec. 4.D) and application of standard statistical significance testing techniques – "the 2 or 3 Standard Deviation Test." These analyses were completed for each year in the four-year life of the eligible list, cumulatively within each candidate pool.

Again, the projected rates at which candidates from each racial/ethnic and gender group would be expected to be selected for further processing are subject to the same caveats noted earlier – these rates may not reflect the actual number of candidates who will pass other steps in the selection process, and might be affected by application of tie-breaking steps applied when more eligible candidates are available than the number of positions.

Evidence of adverse impact under the four-fifths rule was determined by dividing the selection (expected processing) rates in each successive year for racial/ethnic and gender groups (Asian, Black, Hispanic, Native American, and females) by the rates for White, or male, candidates. The fourth-fifths rule is considered to reflect evidence of adverse impact if the selection rate for the minority/female group is less than 80% (4/5ths) the rate for the White, or male, group. This is also called the adverse impact (AI) ratio.

The standard deviation (SD) test uses statistical significance testing to compare the difference between selection rates (e.g., the Black selection rate versus the White selection rate for a given group of candidates), and the result is expressed in units of standard deviation. This difference for any single comparison is assessed in terms of the magnitude of such differences that might be expected to result from pure chance alone. A difference of 2 SD units indicates an approximate 95% probability that the observed difference between the two groups' selection rates exceeds the range of differences that would be expected to result from chance alone. A difference of 3 SD units indicates a greater than 99% probability that the difference exceeds that expected to result from chance alone. Such results provide another means for evaluating evidence of adverse impact.

Results of the adverse impact analyses conducted for candidates' Final Adjusted Score values are shown in Table 26. Each row in Part "a" of the table shows the assumed number of 600 job openings to be filled during each year of the eligibility list, the total number of candidates who remain available for consideration during each successive year, the lowest Adjusted Final Score of those expected to be processed during each successive year, the cumulative selection rate (per year, and within candidate pool), and the AI ratios for Black-White, Hispanic-White, Asian-White, Native American-White, and Male-Female comparisons. Part "b" of the table reports the number of SD units resulting from statistical significance tests for the same groups of candidates reflected in Part "a" of the table.

Table 26. Adverse Impact Analysis

a. Four-fifths Rule

Source	*No. Positions Open*	*No. Candi-dates*	*Minimum Score Obtained*	*Cum. No. Selected to Proceed*	*B-W AI Ratio*	*H-W AI Ratio*	*A-W AI Ratio*	*NA-W AI Ratio*	*F-M AI Ratio*
Internal Pool (N) Year 1	600	873	72	869	0.983	1.002	1.002	1.002	1.005
Open Pool (N)									
Year 2	600	41358	101	2160	0.912	0.909	1.087	0.625	0.809
Year 3	600	39198	99	4708	1.061	0.982	1.167	0.753	0.885
Year 4	600	36650	97	8548	1.037	0.993	1.188	0.801	0.917
Open pool cut score	--	41358	70	40426	0.974	0.985	0.979	0.968	0.999

Note: Cumulative number selected is within pool; AI ratio is based on cumulative pass rates.

b. Standard Deviation Test

Source	*B-W SD's*	*H-W SD's*	*A-W SD's*	*NA-W SD's*	*F-M SD's*
Internal Pool (N) Year 1	-2.26	0.71	0.31	0.07	0.640
Open Pool (N)					
Year 2	-1.64	-1.80	0.73	-1.19	-1.910
Year 3	1.67	-0.52	2.06	-1.16	-1.760
Year 4	1.44	-0.27	3.29	-1.34	-1.810
Open pool cut score	-14.01	-9.12	-5.91	-3.62	-0.270
No. Candidates					
Internal pool	155	226	42	2	80
Open pool	8049	9377	1272	177	1877

Note: No. white candidates = 448 internal pool and 22483 open pool; no. male candidates =793 internal pool and 39481 open pool.

As shown in the tables, no evidence of four-fifths rule adverse impact was observed for any minority or female candidate group across the full four-year life of the eligible list. Only for the very small group of Native American candidates did the four-fifths rule fall below the 80% level during the second and third years of anticipated candidate processing; although rising above the 80% level when processing results were projected over the full four-year life of the eligible list.

The SD test did not yield negative values exceeding 3 SDs in any instance, and exceeded 2 SDs only once – in Year 1 the difference between Black and White internal pool candidates was -2.26 SDs. However, the corresponding AI ratio for this comparison was 0.983, indicating that the selection rates were nearly equivalent for the two groups.

In addition, the SD test showed no evidence of adverse impact for Native American candidates during any years of processing from the eligible list, despite the four-fifths rule results cited above for the second and third years. The positive SD test values exceeding 2 SDs for the Asian-

White comparisons indicated a higher rate of success for Asian (minority) candidates than Whites during Years 3 and 4.

Further, the above described four-year AI and SD test results would still hold for the open pool if the eligible list were to reach down one point lower to a Final Adjusted score of 96, adding 2,340 candidates to the eligible list (a total of 10,888 open pool candidates selected to be processed). That is, all AI ratios would exceed .80 and all SD tests would yield values less than 2 SD units at a Final Adjusted Score of 96.

The Year-1 results for the internal candidate pool, in effect, represent an adverse impact analysis of the cut score for internal candidates since all internal candidates who achieved a passing score would be selected for further processing during this year. With respect to the open pool candidates, the minimum passing score is less relevant, because selection from the eligible list will occur in the upper range of scores (i.e., Final Adjusted Scores of 97 and higher) and the list is expected to expire well before reaching the cut score of 70.

Nevertheless, adverse impact analysis results for the cut score are presented for the open pool in the lower portions of Table 26, Parts "a" and "b." The results indicate no evidence of four-fifths rule adverse impact for racial/ethnic minority or female candidates at the cut score (all AI ratios exceed 0.96). The SD test values for racial ethnic minorities exceed 3 SDs, despite the very similar selection rates among racial/ethnic and gender groups. However, as noted above, the range of scores within which further processing will take place during the life of the eligible list shows no evidence of adverse impact by the SD test.

In summary, there was no evidence that adverse impact would result in significant differences in the selection rates for any minority or female candidate groups from use of Final Adjusted Scores to select open pool candidates over the four-year life of the eligibility list.

Summary

A recommended method for scoring and use of the Firefighter CBT was established prior to its administration to candidates. The method was based on studies of Firefighter incumbents to determine the weighting of test components and derive a minimum passing score. Scoring formulas were derived for converting CBT scores to a 100-point integer scale with a passing score of 70, to which mandated bonus points would be added to yield a final CBT score.

The CBT was administered to 42,231 candidates between March 14 and August 1, 2012 in 15 testing locations in New York City and outlying areas. The sessions were closely supervised by PSI and DCAS proctors to maintain the security of the test and prevent cheating. Four alternate equivalent forms of the CBT were administered to candidates to further ensure test security.

After the completion of candidate testing, analyses were conducted to confirm the accuracy of the scoring key. A Test Validation Board identified modifications to the scoring key and these were implemented to produce candidates' final CBT scores, including bonus points.

Projections of candidate selection rates were made in light of the City's rules and procedures for creating eligibility lists, ranking candidates and selecting approximately 2,400 per year for consideration to fill 600 academy class openings. A total of 9,417 candidates (22.3%) were projected to be selected for consideration to be processed from the eligibility lists over 4 years to fill 2,400 positions. Further analyses of the projected selection rates by race/ethnicity and gender indicated there would be no adverse impact over the anticipated 4-year life of the open competitive pool eligibility list.

REFERENCES

AERA, APA, NCME (1999). *Standards for Educational and Psychological Testing*. AERA, Washington DC.

DCAS (2007). *Job analysis report: Firefighter, Fire Department of New York City: Examination Number 6019*.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978). Uniform Guidelines on Employee Selection Procedures. Washington, DC: Department of Labor.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations*. San Francisco, CA: Jossey-Bass.

Green, S.B., & Veres, J. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology.* Vol. 5, pp. 7-61.

Hunter, J, & Hunter, R. (1984). Validity and Utility of Alternative Predictors of Job Performance. *Psychological Bulletin,* 96 (1).

Nichols, David P. (1998). Choosing an intraclass correlation coefficient. Retrieved from http://support.spss.com/ProductsExt/Statistics/Documentation/Statistics/articles/whichicc.htm.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures, 4th Edition.* Bowling Green, OH.

US Equal Employment Opportunity Commission, US Civil Service Commission, US Department of Justice, & US Department of Labor (1978). *Uniform Guidelines on Employee Selection Procedures*. Federal Register, 43(166), 38295 – 38309